



DTIC FILE COPY



PROCEEDINGS

AD-A215 179

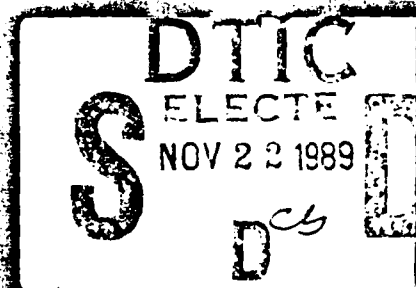
30th Annual Conference of the

MILITARY TESTING ASSOCIATION

Arlington, Virginia

27 November - 2 December 1988

Edited by
Arthur C. F. Gilbert



DISTRIBUTION STATEMENT A

Approved for public release
Distribution Unlimited

Coordinated by the U.S. Army Research Institute
for the Behavioral and Social Sciences

**Best
Available
Copy**



PROCEEDINGS

30th Annual Conference of the

MILITARY TESTING ASSOCIATION

Arlington, Virginia

27 November - 2 December 1988

Edited by
Arthur C. F. Gilbert

Accession For	
NTIS CRA&I	<input checked="" type="checkbox"/>
DTIC TAB	<input type="checkbox"/>
Unannounced	<input type="checkbox"/>
Justification	
By <i>Dr. Gilbert</i>	
By <i>11/22/89</i>	
Distribution <i>ARI</i>	
Availability Codes	
Dist	Avail and/or Special
<i>A-1</i>	



Coordinated by the U.S. Army Research Institute
for the Behavioral and Social Sciences

89 11 21 030

PROCEEDINGS
30TH ANNUAL CONFERENCE
OF THE
MILITARY TESTING ASSOCIATION

Coordinated by the
U.S. Army Research Institute
for the Behavioral and Social Sciences

Arlington, Virginia
27 November - 2 December 1988

CONFERENCE COMMITTEE

30TH ANNUAL CONFERENCE OF THE MILITARY TESTING ASSOCIATION (MTA)

Arlington, Virginia

27 November - 2 December 1988

President:	Colonel Jon W. Blades
Conference Coordinator:	Charlie Holman
Chair, Program Committee:	Arthur C. F. Gilbert
Secretary:	Robert F. Holz
Membership:	Arthur. C. F. Gilbert
Registration:	Charlie Holman
Finance:	Charlie Holman
Social:	Clinton B. Walker
Editor, Proceedings:	Arthur C. F. Gilbert

ACKNOWLEDGEMENTS

The MTA Conference Committee wishes to express, on behalf of the members of the Association, its appreciation for the voluntary and excellent service provided by the following persons in support of this year's conference.

Administration

Tracye Julien
Jessie McVean
Abby von Till

Audio-Visual

Major Brian Whiting

Graphics

Sheila Bombardiere
John H. Nagel
Norma C. Rollins

Hospitality

Mark Czarnolewski
Scott E. Graham
Melvin J. Kimmel
Richard E. Maisano
Arthur Marcus
Franklin L. Moses
Nigel Nicholson
Joel M. Savell
Guy L. Siebold
Alfred L. Smith, Jr.
Edwin R. Smootz
Uldi Shvern
John E. Stewart, II
Leonard A. White

Logistics

Major Richard Cline
Karlton Kinzer
Major Brian Whiting

Program

Sheila Bombardiere
Arlene Crow
Mary C. Caswell
Sheldon A. Glazer
William Harlow
H. Joanna Thayer
Bonnie Tolbert

Registration

Major Richard Cline
Laurel Oliver
Ok-Choon Park
Betty Shelly
Annie D. Stark
Donna Thurber
Bonnie Tolbert
Paul Twohig
Major Brian Whiting

Session Chairs

Jane M. Arabian
Michael E. Benedict
Barbara A. Black
Scott E. Graham
Christine R. Hartel
Arthur Marcus
David Promisel
Michael G. Rumsey
Uldi Shevern
Zita M. Simutis
John E. Stewart, II
Nora Stewart
Robert A. Sulzen
Patrick J. Whitmarsh
Robert A. Wisher

FOREWORD

These Proceedings of the 30th Annual Conference of the Military Testing Association document the presentations given at paper and panel sessions during the Conference. The papers represent a broad range of topics by contributors from the military, industrial, and educational communities both foreign and domestic. It should be noted that the papers reflect the opinions of the authors and do not necessarily reflect the official policy of any institution, government, or armed services agency.

CONTENTS

1988 MTA CONFERENCE COMMITTEE	i
ACKNOWLEDGEMENTS	iii
FOREWORD	v
TABLE OF CONTENTS	vii
OPENING SESSION	1

HUMAN FACTORS — PAPER PRESENTATIONS

MANPRINT AIDS TO ASSESS SOLDIER QUALITY AND QUANTITY. Richard D. Herring and Lawrence H. O'Brien	3
MANPRINT EVALUATION OF THE THEATER ARMY MEDICAL MANAGEMENT INFORMATION SYSTEM. Norman D. Smith and John R. Tiffany	9
TESTING AQUILA, THE ARMY'S REMOTELY PILOTED VEHICLE: LESSONS LEARNED. Edwin R. Smootz and Nigel R. Nicholson	15
ESTIMATING MAINTENANCE MANPOWER REQUIREMENTS FOR THE AQUILA REMOTELY PILOTED VEHICLE. John E. Stewart, II	21
NDI: A REAL SHORTCUT TO SYSTEM TESTING? John L. Miles, Jr. and Jonathan D. Kaplan	27
ERROR IS ALSO A HUMAN FACTOR: TASK DIFFICULTY IN CONTEXT. Carl W. Lickteig and Robert S. Du Bois	33
THE RESULTS OF A SEARCH FOR COGNITIVE SKILLS. John A. Modrick	39
APPLICATIONS OF THE COGNITIVE REQUIREMENTS MODEL. Paul G. Rossmeissl and Jan Charles	45
SOLDIER FEEDBACK FOR BETTER PRODUCTS. Barbara A. Jezior, Lawrence E. Symington, Charles A. Greene, and Annette M. Salvato	51
HOW MUCH DO SOLDIERS SLEEP? Sally J. Van Nostrand	57

HUMAN FACTORS — PANEL PRESENTATIONS

PANEL - TESTING VISION, PERFORMANCE, AND SUBJECTIVE SYMPTOMATOLOGY IN REDUCED OXYGEN ENVIRONMENTS. Richard L. Burse, Sc.D. (Chair)	63
VISUAL SENSITIVITIES UNDER REDUCED OXYGEN. S. M. Luria, Nancy Morris and Alan Cymerman	64

COGNITIVE AND MOTOR PERFORMANCE UNDER REDUCED OXYGEN. Christine L. Schlichting, Douglas R. Knight, and Alan Cymerman	70
COGNITIVE PERFORMANCE AND MOOD STATES IN 13-21% OXYGEN ENVIRONMENTS. Barbara L. Shukitt and Louis E. Banderet	76
ALTITUDE ILLNESS SYMPTOMATOLOGY IN 13, 17, AND 21% OXYGEN ENVIRONMENTS. Richard L. Burse, Charles S. Fulco, Allen Cymerman, and Douglas R. Knight	82

LEADERSHIP — PAPER PRESENTATIONS

HOW SMALL UNIT COHESION AFFECTS PERFORMANCE. Guy L. Siebold	87
THE RELATIONSHIP BETWEEN LEADERSHIP COMPETENCY RATINGS AND PLATOON COHESION. Dennis R. Kelly	93
COMBAT READINESS: PREPARING SOLDIERS FOR THE STRESS OF BATTLE. Major W. R. Wild	99
MEASURING LEADER PERFORMANCE IN SIMULATED COMBAT IN THE FIELD. Paul T. Twohig and Trueman R. Tremble	105
ATTITUDES OF GFAF RESERVISTS TOWARDS RESERVE DUTY TRAINING. Sibylle B. Schambach	111
ACTUAL CONTRIBUTIONS OF MILITARY PSYCHOLOGY TO MANPOWER MANAGEMENT. Friedrich W. Steege	117
LEADERSHIP ISSUES IN GENDER INTEGRATION. Major R. A. V. Dickenson	123
WHAT IS THIS THING CALLED CHARISMA? Leanne E. Atwater, Ph.D., Robert Penn, Ph.D., and Linda Rucker, M.A.	129
THE EFFECTS OF INDIVIDUAL EXPERIENCE AND INTELLIGENCE ON DECISION PERFORMANCE. Jody C. Locklear, Charles G. Powell, and Fred E. Fiedler	135
THE EFFECT OF STRESS ON THE PERFORMANCE OF CREATIVE AND INTELLIGENT LEADERS. Fred E. Fiedler and Frederick W. Gibson	141
COST-EFFECTIVE RETENTION OF WEST POINT OFFICERS. Hyder Lakhani and Rashmi Lal	147
A NON-OBTRUSIVE METHOD OF EVALUATING TACTICAL DECISIONMAKING IN THE FIELD. Marvin L. Thorsden, Gary A. Klein, Rex R. Michel, and Major Edward W. Sullivan	153
MEASURING DIVISION LEVEL COMMAND AND CONTROL PERFORMANCE. Dr. Lloyd M. Crumley	159

SIMULATION-BASED C ³ TESTING FOR ARMOR PLATOON LEADERS. Robert S. Du Bois, and Carl W. Lickteig	165
ARMY COMMISSIONED AND NONCOMMISSIONED OFFICER LEADER REQUIREMENTS. Alma G. Steinberg and Julia A. Leaman	171
WHAT DO ARMY SERGEANTS MAJOR IN STAFF POSITIONS DO? Gilbert L. Neal, James D. Moreland, John LaVerne, and F. Edward Saia	177
STAFF OFFICER CHARACTERISTICS CONTRIBUTING TO EFFECTIVE TACTICAL DECISION MAKING. William D. Sprenger, Jon J. Failesen, and Sharon L. Riedel	183
VIETNAM: LASTING EFFECTS ON CONFIDENCE TOWARD MILITARY LEADERS. Frank J. Ricotta, Jr., Charles N. Weaver, and Michael D. Matthews	189

LEADERSHIP — PANEL PRESENTATIONS

PANEL - PREDICTING LEADERSHIP AT THE SERVICE ACADEMIES AND BEYOND. Dr. Robert F. Priest (Chair)	193
ASSESSING LEADERSHIP POTENTIAL AT THE NAVAL ACADEMY WITH A BIOGRAPHICAL MEASURE. Lawrence J. Stricker	194
PREDICTING LEADERSHIP AT THE SERVICE ACADEMIES AND BEYOND. Craig J. Russell and Karl W. Kuhnert	200
THE U. S. COAST GUARD ACADEMY CLASS OF 1986: BIOGRAPHICAL PREDICTORS OF SUCCESS. Earl H. Potter and Robert R. Albright	204
PREDICTING LEADERSHIP AT THE SERVICE ACADEMIES AND BEYOND: DISCUSSANT REMARKS. Martin F. Wiskoff (Discussant)	210

MANPOWER TRENDS — PANEL PRESENTATIONS

MEDICAL STANDARDS FOR ENLISTMENT AND THE QUALIFIED MILITARY AVAILABLE POPULATION. Michael T. Laurence	216
CHILDREN OF MILITARY: FORGOTTEN COMBAT MULTIPLIERS? Michael E. Freville, Ed.D. and Rose Webb Brooks, M.A.	222
THE IMPACT OF SPOUSE EMPLOYMENT ON MILITARY MANPOWER. Paul A. Gade, Newell Kent Eaton and Royce Bauman	225
CONTRIBUTIONS OF SPOUSE RELATED FACTORS AFFECTING REENLISTMENT OF ENLISTED PERSONNEL. Alfred L. Smith, Jr.	231

FAMILY DEMOGRAPHIC COHORTS: MAKERS/BREAKERS OF COHORT UNIT COHESION AND READINESS. Joel M. Teitelbaum, Ph.D. and LTC T. Paul Furukawa, Ph.D.	237
GETTING BETTER RESPONSES FROM SPOUSES. Morris Peterson, Ph.D., Susan Kerner-Hoeg, and Emily Cato	244

OCCUPATIONAL ANALYSIS — PAPER PRESENTATIONS

A FUNCTIONAL EVALUATION OF A MODEL ARMY PERSONNEL ADMINISTRATION CENTER (PAC) ORGANIZATION. Raymond O. Waldkoetter and Phillip L. Vandivier ...	250
A JOB ANALYSIS OF THE ROLE OF A POLICE OFFICER. Sheldon H. Geller, Marijane Terry, and Fred Shaw	256
THE OCCUPATIONAL RESEARCH DATA BANK: A KEY TO MPT SUPPORT. 1Lt Kathleen M. Longmire and Lt Col Lawrence O. Short.....	262
DEVELOPING A TOTAL FORCE OCCUPATIONAL SPECIFICATION FOR THE CANADIAN FORCES. LCDr Dominique D. Benoit and Mr. G. Jeffrey Higgs.....	263
AN APPLICATION OF KSA ANALYSIS TO SELECTION AND TRAINING DECISIONS. Janet George Irvin and Janet H. Blunt	274
AFFORDABLE AND CREDIBLE PROCEDURES FOR DETERMINING OCCUPATIONAL LEARNING DIFFICULTY. Phillip A. Davis, SQDNLDR, RAAF	280
JOB INCUMBENT PERFORMANCE REPORTED BY ABSOLUTE FREQUENCY - A FURTHER EXAMINATION. Lawrence A. Goldman, Ph.D.	286
CODAP: ORGANIZING APPLICATIONS AND R&D OUTSIDE OF THE U.S. MILITARY. Michael R. Staley and Johnny J. Weissmuller	292
A COMPARISON OF METHODOLOGIES FOR GROUPING LARGE NUMBERS OF OCCUPATIONS. Julie Rheinstein, Donald E. McCauley, Jr., and Brian S. O'Leary	298

OCCUPATIONAL ANALYSIS — PANEL PRESENTATIONS

PANEL - OCCUPATIONAL ANALYSIS: PRESENT AND FUTURE. Dr. Hendrick W. Ruck (Chair)	304
OCCUPATIONAL ANALYSIS: THE PRESENT. J. S. Tartell	305
OCCUPATIONAL ANALYSIS: THE FUTURE. Dr. Hendrick W. Ruck	310
Dr. Hendrick W. Ruck and J. S. Tartell (Discussants)	

PANEL - DEVELOPMENT AND DESIGN OF THE NAVY OFFICER OCCUPATIONAL DATABASE. Captain Edward L. Naro, USN and Dr. Janet M. Treichel (Co-Chairs)	316
DEVELOPMENT AND DESIGN OF THE NAVY'S OFFICER OCCUPATIONAL TASK ANALYSIS PROGRAM. Captain Edward L. Naro, USN	317
THE NAVY OFFICER SURVEY INSTRUMENT (OSI). LT Susan J. Fiorino, USN	323
THE NAVY MEDICAL COMMUNITY OFFICER OCCUPATIONAL TASK ANALYSIS PROGRAM (NOTAP) SURVEY. LCDR M. Ellen Quisenberry	328
PANEL - NEW ASCII CODAP TECHNOLOGY: MANPOWER, PERSONNEL, TRAINING APPLICATIONS. Dr. Walter E. Driskill (Chair)	334
INTRODUCTION TO OPERATIONAL ASCII CODAP: AN OVERVIEW. Johnny J. Weissmuller, Joseph S. Tartell, and William J. Phalen	335
ASCII CODAP PROGRAMS FOR SELECTING AND INTERPRETING TASK CLUSTERS. William J. Phalen, Michael R. Staley, and Jimmy L. Mitchell	341
OPERATIONAL TESTING OF ASCII CODAP JOB AND TASK CLUSTERING METHODOLOGIES. Jimmy L. Mitchell, William J. Phalen, William Haynes, and Darryl Hand	347
ASCII CODAP AND MANPOWER-PERSONNEL-TRAINING (MPT) TECHNOLOGIES. R. Bruce Gould, Hendrick W. Ruck, Walter E. Driskill, and Jay S. Tartell	353
Lt Col Frank C. Gentner (Discussant)	
PANEL - NEW TECHNOLOGIES FOR DEVELOPING AUTOMATED DATA-BASED SPECIALTY KNOWLEDGE OUTLINES. Paul P. Stanley II (Chair)	359
AUTOMATED TEST OUTLINE DEVELOPMENT RESEARCH FINDINGS. Johnny J. Weissmuller, Martin J. Dittmar, and William J. Phalen	360
DEVELOPMENT OF AUTOMATED DATA-BASED SPECIALTY KNOWLEDGE TEST OUTLINES: CURRENT PROCEDURES. 1Lt Kathleen M. Longmire, William J. Phalen, Johnny J. Weissmuller, and Martin J. Dittmar	366
AUTOMATED SPECIALTY KNOWLEDGE TEST OUTLINE PROCEDURE: A MANAGEMENT PERSPECTIVE. Paul P. Stanley II, Rondald C. Baker, and Joseph S. Tartell	372
AUTOMATED SPECIALTY KNOWLEDGE TEST OUTLINE PROCEDURES: A DEVELOPMENT TEAM PERSPECTIVE. John E. Williams, 1Lt, USAF, Wendy L. Sotello, 1Lt, USAF, and Paul P. Stanley II	378

TESTING — PAPER PRESENTATIONS

THE FUTURE OF ITEM ANALYSIS. Howard Wainer	384
THE BUROS INSTITUTE OF MENTAL MEASUREMENTS IN 1990s. Barbara S. Plake	390
ITEM BANKING IN SKILL QUALIFICATION TEST (SQT) DEVELOPMENT. Allan L. Pettie	395
TRAINING TEST ITEM DEVELOPERS: A NEW APPROACH. Harvey Rosenbaum	399
COMPUTERIZED ITEM BANKING AND TEST ASSEMBLY. Lawrence S. Buck	404
TEST-ITEM READABILITY: A FINAL REPORT. R. Eric Duncan, Captain, USAF	410
PREDICTION OF SENSITIVE COMPARTMENTED INFORMATION (SCI) ACCESS USING PERSONALITY TESTS. LeRoy A. Stone, Ph.D., ABPP. ABFP	416
A MICROCOMPUTER TEST BATTERY: NORMATIVE DATA AND SENSITIVITY TO MILITARY STRESSORS. Robert S. Kennedy, Dennis R. Baltzley, and Mary K. Osteen	422
A BIODATA INSTRUMENT FOR CIVIL SERVICE EXAMINATIONS: INITIAL VALIDATION. Jay A. Gandy, Alice N. Outerbridge, and James C. Sharf	428
SYNTHETIC VALIDATION PROCEDURES FOR IDENTIFYING SELECTION COMPOSITES AND CUT SCORES. Jane M. Arabian, Jeffrey J. McHenry, and Lauress L. Wise	434
PSYCHOMETRIC PROPERTIES OF THREE ADDITION TASKS WITH DIFFERENT RESPONSE REQUIREMENTS. L. E. Banderet, Ph.D., B. L. Shukitt, B.A., SSG Michael A. Walthers, Robert S. Kennedy, Ph.D., Alvah C. Bittner, Jr., Ph.D., and Gary G. Kay, Ph.D.	440
COMPUTERIZED TESTING (CAT) IN THE GERMAN FEDERAL ARMED FORCES (GFAF). Wolfgang Wildgrube	446
FUNCTIONALITY AND ARCHITECTURE OF THE GFAF COMPUTERIZED ADAPTIVE TESTING SYSTEM. Dr. Michael Habon	452
COMPUTERIZED ADAPTIVE TESTING: THE CAT-ASVAB PROGRAM. W. A. Sands	458
FUTURE TESTS - DESIGN FOR VALIDATION IN TEN NAVY SCHOOLS. John H. Wolfe	463
VALIDATION OF THE AIR FORCE RESERVE OFFICER TRAINING CORPS SELECTION SYSTEM. Linda R. Elliot+	469
THE RELATIONSHIP BETWEEN APTITUDE AND ARMY OFFICER PERFORMANCE. Dianne C. Brown	475

DEVELOPMENT OF A NAVAL OFFICER SELECTION TEST. Lieutenant (N) Alan C. Okros	481
COMPLEX COGNITIVE ASSESSMENT BATTERY: PERFORMANCE, DEMOGRAPHIC AND ATTITUDINAL ASPECTS. William D. Sprenger, Ph.D. and Jon T. Fallesen, Ph.D.	486
APPLICATION OF PERFORMANCE PROTOCOL ANALYSIS IN MILITARY TESTING. Michael G. Samet and Christine Hartel	492
PREDICTING TANK GUNNERY PERFORMANCE FROM CREWMEMBERS' EXPERIENCE AND COGNITIVE ABILITY. R. Gene Hoffman and David A. Campshure	498
THE UNIFIED TRI-SERVICE COGNITIVE PERFORMANCE ASSESSMENT BATTERY. G. Rufus Sessions, David R. Thorne, and Samuel L. Moise, Jr., and Frederick W. Hegge	504
RE-EVALUATING PRACTICAL PERFORMANCE ASSESSMENT STRATEGIES IN THE CANADIAN NAVY. Lcdr Edward G. Barnett, CF	510
A FULLY AUTOMATED MEMORY AND SEARCH TASK. Charles A. Salter, Laurie S. Lester, Heather Dragsbaek, Richard D. Popper, and Edward Hirsch	515
ARMED FORCES VOCATIONAL APTITUDE BATTERY AND VEHICLE IDENTIFICATION PERFORMANCE RELATIONSHIPS. Otto Heuckeroth and Norman D. Smith	521
NEW ASSESSMENT FOR SHORT-SERVICE VOLUNTEERS. Axel R. Kaiser	527
METHODOICAL AND ORGANIZATIONAL DEVELOPMENT OF SELECTION: PROGRESS AND RESULTS. Albert H. Melter	532
TESTING U.S. ARMY WAR COLLEGE STUDENTS' WRITING ABILITIES. Colonel Robert J. Davis and Professor Jim Hanlon	538
DEVELOPMENT OF A NEW PSEUDO AFQT TO DETECT POSSIBLE COMPROMISE. Thomas W. Watson, Steven W. Hoffer, Sgt, USAF, and Malcolm James Ree	542
PRE-OPERATIONAL VALIDATION OF NEW ARMY FLIGHT APTITUDE SELECTION INSTRUMENTS. D. Michael McAnulty	548
RETAKING THE U.S. NAVY AND MARINE CORPS AVIATION SELECTION TEST BATTERY. Annette G. Baisden and LCDR F. Douglas Holcombe	554
CROSS-VALIDATION OF AN EXPERIMENTAL PILOT SELECTION AND CLASSIFICATION TEST BATTERY. Thomas R. Carretta, Ph.D.	559
CURRENT DIRECTIONS IN THE AIR FORCE PILOT SELECTION AND CLASSIFICATION RESEARCH. Frederick M. Siem, Dwight C. Hageman, and Theresa A. Mercatante	565
ISSUES IN JOB SAMPLE TESTING. Herbert George Baker, Ph.D. and Gerald J. Laabs, Ph.D.	571

A SYSTEMS APPROACH TO THE SYNTHESIS OF MEASURES OF EFFECTIVENESS. Edward Connelly	576
MOMENT TO MOMENT PERFORMANCE ASSESSMENT OF MILITARY SYSTEMS. Edward Connelly	582
DIAGNOSING TRAINEE LEARNING DIFFICULTIES THROUGH LEVELS OF ACHIEVEMENT ANALYSIS. George M. Usov, Ph.D.	588
MEASURES OF EFFECTIVENESS: INCREASING THEIR SENSITIVITY TO PERSONNEL CHARACTERISTICS. Mark V. Czarnolewski and John L. Miles, Jr.	591
TOWARD A DIAGNOSTIC OPERATOR PERFORMANCE ASSESSMENT SCHEME. Dr. Laurel Allender and Mr. Bryan E. Brett	597
A COMPREHENSIVE METHOD FOR EVALUATION OF CRITICAL MILITARY TRAINING PROGRAMS. Jeffrey A. Cantor and C. Lee Walker	602
CIVILIAN TEST EXPERTS HELP IMPROVE SKILL QUALIFICATION TESTS. Paul R. Vaughan and Clay V. Britvain	608
TESTING -- PANEL PRESENTATIONS	
PANEL - THE U.S. MARINE CORPS INFANTRYMAN JOB PERFORMANCE MEASUREMENT PROJECT. Milton D. Maier (Chair)	614
TASK SELECTION AND TEST DEVELOPMENT FOR THE INFANTRYMAN JPM PROJECT. Daniel B. Felker and Charles W. Harnest	615
QUALITY CONTROL PROCEDURES AND INTERRATER RELIABILITY RESULTS. Jennifer L. Crafts, Edmund C. Bowler, and David W. Rivkin	621
QUALITY CONTROL PROCEDURES AND RELIABILITY RESULTS. Andrew M. Rose and Jennifer L. Crafts ..	627
VALIDITY RESULTS FROM THE MARINE CORPS JOB PERFORMANCE MEASUREMENT PROJECT. Paul W. Mayberry	633
Richard Harris (Discusant)	
PANEL: EVALUATING PROJECT A TESTS FOR SELECTING GUNNERS. Clinton B. Walker (Chair)	640
SELECTING SOLDIERS FOR THE EXCELLENCE IN ARMOR PROGRAM. Scott E. Graham	641
TESTING PSYCHOMOTOR AND SPATIAL ABILITIES TO IMPROVE TOW GUNNER SELECTION. Elizabeth D. Smith and Martin R. Walker	647

EVALUATING PSYCHOMOTOR AND SPATIAL TESTS FOR SELECTING AIR DEFENSE GUNNERS. Ilene F. Gast and David M. Johnson	652
Jeffrey J. McHenry (Discussant)	

TRAINING — PAPER PRESENTATIONS

MAINTAINING COURSE CURRENCY IN A CHANGING ENVIRONMENT. Mollie J. Tucker	659
NAVY CLASSROOM TRAINING: A STATUS REPORT. Barbara Taylor, John Ellis, and Robyn Baldwin	663
OVERSEAS MAIL TIME FOR CORRESPONDENCE INSTRUCTION. Grover E. Diehl	669
INTERACTIVE VIDEO: R & D AND A PRACTICAL APPLICATION IN THE BRITISH ARMY. Colonel Donald H. Oxley and Major Dennis Quilter	675
SIMULATION, CBT & TESTING COMPRISE, CONFOUND OR CAMP? Cdr. Robert H. Kerr, CF, Maj. Marilyn Hoggard, CF, and Capt. Bruce D. Hyland, CF	681
APPLICATIONS OF SIMULATION AND WARGAMING TO TRAINING. Franklin L. Moses and Earl A. Alluisi	686
COMPUTER BASED TRAINING (CBT) - A NAVAL TRAINING EVALUATION. Lieutenant Commander Peter J. Ross, Royal Navy, Lieutenant Commander S. L. Latimer, Royal Australian Navy, and Lieutenant Commander A. R. Jones, Royal Navy	692
TRAINING FOR A FLYING EMERGENCY - A TASK ANALYSIS OF ENGINE-OFF-LANDINGS. Amanda J. W. Feggetter and Heather M. McIntyre	698
HELICOPTER INSTRUMENT PROCEDURES TRAINER (HIPT): AN EVALUATION OF TRAINING EFFECT. Captain I. E. (Ed) Wiebe and Captain G. A. (Greg) Reiser	704
PERFORMANCE EVALUATION IN TACTICAL TRAINERS. John A. Modrick and Thomas A. Plocher	710
EFFECTIVENESS OF AN ADVANCED INDIVIDUAL COMBAT ARMS TRAINER (AICAT). Brigadier General Charles R. White, Captain Jarean L. Carson, Captain Keith R. Wynkoop, Lieutenant Von M. Cameron, and Clifford A. Butzin, Ph.D.	716
JOINT SERVICE TRAINING REQUIREMENTS DECISION SUPPORT SYSTEM. Dr. H. Barbara Sorensen and Ms Marlene R. Laskowski	722
USAF INTEGRATED MANPOWER, PERSONNEL AND COMPREHENSIVE TRAINING & SAFETY (IMPACTS) PROGRAM. Elaine Howell, Major, USAF	727
USAF AERONAUTICAL SYSTEMS DIVISION'S MODEL MANPOWER, PERSONNEL AND TRAINING ORGANIZATION--AN UPDATE. Lt Col Frank C. Gentner	731

ESTABLISHING A RELATIONSHIP BETWEEN TRAINING RESOURCES EXPENDITURES AND UNIT PERFORMANCE. Brian J. Bush	737
TRAINING SYSTEMS ANALYST: THE CHANGING ROLE OF THE BEHAVIORAL PSYCHOLOGIST. Major Conrad G. Bills, Lt Nancy J. Fakult, Lt Z. Nagin Ahmed, and James E. Brown	740
A STUDY OF BEHAVIOR MODELING IN MANAGEMENT TRAINING. Dr. Phyllis Peters Marson	744
MODELING THE COSTS AND BENEFITS OF ALTERNATIVE TRAINING INTERVENTIONS. Michael D. Mumford, Joseph L. Weeks, Francis D. Harding, and Edwin A. Fleishman	749
THE EFFECTS OF A MATHEMATICS REFRESHER COURSE ON SCHOOL ATTRITION. Gary R. Bunde	755
ACCURACY AND ADAPTABILITY: AN INVESTIGATION OF STANDARD DISTANCE ESTIMATION PROCEDURES. Mark A. Guadagnoli, Gene W. Fober, Pamela M. Terry, and William R. Harden	761
INTERMEDIATE FORWARD TEST EQUIPMENT TRAINING EFFECTIVENESS ANALYSIS. Kathy L. Nau and Gary G. Sarli	767
DEVELOPMENT OF THE ADVANCED ON-THE-JOB TRAINING SYSTEM: LESSONS LEARNED. Bernie Marrero, Ph.D.	773
TRAINING -- PANEL PRESENTATIONS	
PANEL - THE AIR FORCE TRAINING DECISIONS SYSTEM: R & D RESULTS. Winston R. Bennett (Chair)	778
OVERVIEW OF TRAINING DECISIONS SYSTEM: RESULTS AND PRODUCTS. David S. Vaughan	779
THE TASK CHARACTERISTICS SUBSYSTEM: ALLOCATING TASK MODULES TO TRAINING SETTINGS. Bruce M. Perrin and J. R. Knight	785
THE FIELD UTILIZATION SUBSYSTEM: JOB AND TRAINING PATTERN SIMULATIONS. J. L. Mitchell and R. M. Yadrick	791
RESOURCE/COST SUBSYSTEM: ESTIMATING TRAINING CAPACITIES AND COSTS. F. H. Rueter and Steve Feldsott	797
INTEGRATION/OPTIMIZATION SUBSYSTEM: AN INTEGRATED MODELING APPROACH. David S. Vaughan and A. John Eschenbrenner	803
DISCUSSION OF THE TDS PROJECT. Hendrick W. Ruck (Discussant)	809

PANEL - TRAINING, RETENTION, AND SUSTAINMENT OF FOREIGN LANGUAGE SKILLS IN THE ARMY. Zita M. Simutis (Chair)	810
A PRELIMINARY INVESTIGATION OF THE RELATIONSHIP BETWEEN ASVAB AND DLAB. Leonard A. White, Lawrence M. Hanser, and Randolph K. Park	811
THE LANGUAGE SKILL CHANGE PROJECT (LSCP). John A. Lett, Jr.	817
DEVELOPMENT AND EVALUATION OF THE STRATEGY INVENTORY FOR LANGUAGE LEARNING. Rebecca L. Oxford, Ph.D.	822
IDENTIFYING PRECURSORS OF SUCCESS IN FOREIGN LANGUAGE. Francis E. O'Mara, Ph.D.	826
NEW TECHNOLOGIES FOR LANGUAGE ASSESSMENT AND SUSTAINMENT. Melissa Holland, Stan Kostyla, Merryanna Swartz, and Joe Psotka	832
PANEL - RESEARCH ON TRAINING WITH SIMULATED NETWORKING TECHNOLOGY (SIMNET). Dr. Jack H. Hiller (Chair)	838
THE SIMNET OPPORTUNITY FOR RESEARCH ON TRAINING AND PERFORMANCE MEASUREMENT. Jack H. Hiller	839
THE SIMNET RESEARCH PROGRAM: OVERVIEW AND PERFORMANCE MEASUREMENT SYSTEM PROGRESS. Richard W. Vestewig, Ph.D.	843
GLOBAL TRAINING STRATEGY FOR THE SIMULATION NETWORK. Jim L. Madden	849
THE SIMNET RESEARCH PROGRAM: AN OVERVIEW. Thomas J. Lubaczewski	855
STRATEGIES FOR UNIT PERFORMANCE MEASUREMENT IN SIMNET. Nancy K. Atwood and William J. Doherty	859
FEEDBACK STRATEGIES FOR SIMNET. Judith J. Nichols and James W. Kerins	864
Dr. Barbara Black and Colonel Samuel Wasaff (Discussants)	
PANEL - ISSUES IN MEASURING UNIT PERFORMANCE AT COMBAT TRAINING CENTERS William J. Doherty (Chair)	868
PRACTICAL SOLUTIONS TO THE CRITERION PROBLEM AT COMBAT TRAINING CENTERS. Jack H. Hiller	869
BRIGADE PERFORMANCE MEASUREMENT SYSTEM. James T. Root	876
ASSESSING LIGHT FORCES AT THE JOINT READINESS TRAINING CENTER. Judith J. Nichols and LTC Howard W. Crawford	882
MEASUREMENT ISSUES AT THE BATTLE COMMAND TRAINING PROGRAM (BCTP). William A. Ross and Karol Girdler	888

USING SIMULATED NETWORKING TECHNOLOGY AS A COMBAT TRAINING CENTER SURROGATE. William J. Doherty and Nancy K. Atwood	894
Colonel Kent Harrison (Discussant)	
PANEL - STANDARDIZATION OF TRAINING MEASUREMENT AT COMBAT TRAINING CENTERS (CTCs). Robert H. Sulzen (Chair)	900
A SYSTEMS APPROACH TO TACTICAL ASSESSMENT. James T. Root	901
MEASUREMENT AT THE JOINT READINESS TRAINING CENTER. Major Jose G. Ventura, Jr.	906
THE U. S. ARMY'S TRENDLINE ANALYSIS PROGRAM. MAJ Joseph R. McLaughlin	911
ARTEP MISSION TRAINING PLAN USE AT THE CTC'S. David M. Atwood, Major, U.S. Army	915
DEVELOPMENT OF OBSERVER CONTROLLER (O/C) GUIDEBOOKS. Robert H. Sulzen	919
PROTOTYPE ELECTRONIC CLIPBOARD SOFTWARE AND HARDWARE FOR THE COMBAT TRAINING CENTERS (CTCs). Patrick J. Whitmarsh	923
Jack Hiller and COL Lee Greene (Discussants)	

CONFERENCE INFORMATION

MINUTES OF THE 1988 STEERING COMMITTEE MEETING	927
STEERING COMMITTEE LIST OF MEMBERS AND ATTENDEES	931
HARRY H. GREER AWARD	933
BY-LAWS OF THE MILITARY TESTING ASSOCIATION	937
AGENCIES REPRESENTED BY MEMBERSHIP ON THE MTA STEERING COMMITTEE	943
CONFERENCE REGISTRANTS	945
INDEX OF AUTHORS AND PANEL MEMBERS	961

30TH ANNUAL CONFERENCE OF THE MILITARY TESTING ASSOCIATION

Arlington, Virginia

28 November 1988

OPENING SESSION

Opening Remarks: Colonel Jon W. Blades, Commander, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, Virginia

Welcome: Dr. Edgar M. Johnson, Technical Director, U.S. Army Research Institute for the Behavioral and Social Sciences and Chief Psychologist of the Army

Keynote Address: LTG A. K. Ono, Deputy Chief of Staff for Personnel, Department of the Army, "Recruiting"

"Other Manpower Trends and Challenges to Research"

RADM Henry McKinney, Commander, U.S. Navy Recruiting Command

LtGen J. I. Hudson, Deputy Chief of Staff for Manpower, U.S. Marine Corps

Lt Gen Thomas J. Hickey, Deputy Chief of Staff for Personnel, Department of the Air Force

Questions and Answers

MANPRINT AIDS TO ASSESS SOLDIER QUALITY AND QUANTITY¹

Richard D. Herring and Lawrence H. O'Brien
Dynamics Research Corporation (DRC)
Wilmington, MA

This paper describes three automated aids for estimating MANPRINT factors for Army weapon systems during of the acquisition process. The Army Research Institute (ARI) is developing these aids under its project to provide improved MANPRINT methods.

The M-CON Aid estimates the maximum manpower that is likely to be available to support a new weapons system. Analysts can use this aid to set manpower constraints, compare these constraints with system requirements, and change constraints after conducting sensitivity analyses.

The P-CON Aid estimates the likely future distribution of key personnel characteristics in particular subpopulations of soldiers. The aid also predicts the levels of human performance that can be expected at various levels of these personnel characteristics. Army analysts and contractors can use this information to set personnel constraints that are based upon performance.

The Personnel-Based System Evaluation (PER-SEVAL) Aid will assist Army analysts in identifying the quality of personnel needed to support a particular contractor's design. The aid predicts task performance as a function of key personnel characteristics, degrades these performance estimates to reflect critical environmental stressors, and estimates system performance by simulating networks of operator tasks and maintenance actions.

M-CON Aid

The M-CON Aid will produce four measures of quantitative manpower constraints. Two of these measures, maximum operator crew size and maximum maintenance manhours, will describe manpower constraints for a single system. The other two measures, total operator manpower requirements and total maintenance manpower requirements, will describe the total pool of manpower available to man the new system. The maintenance manpower constraints are broken out by paygrade and Military Occupational Specialty (MOS) at each maintenance level (e.g., organizational, direct support, and general support).

Assumptions

The M-CON Aid focuses on major weapon systems that shoot; however, the basic logic of the M-CON Aid should be applicable to other systems as well. It is important to remember that the purpose of M-CON Aid is to identify the maximum manpower that is most likely to be available to support a new weapon system; not to identify the manpower required. (Another MANPRINT methods aid will identify manpower requirements.) The M-CON Aid manpower calculations are based on Table of Organization and Equipment (TOE) units in a peacetime environment since they impose the most restrictive constraints. Support personnel that cannot be directly tied to a maintenance or operator position are not considered.

¹ Development of these aids is being conducted under Contract # MDA 903-86-R-0140.

Method Overview

Figure 1 provides an overview of the method used to estimate manpower constraints. The M-CON Aid estimates manpower constraints by identifying the manpower that will be available to man the new system. Since total Army end strength is fixed, the manpower slots needed to support a new system must be drawn from existing systems.

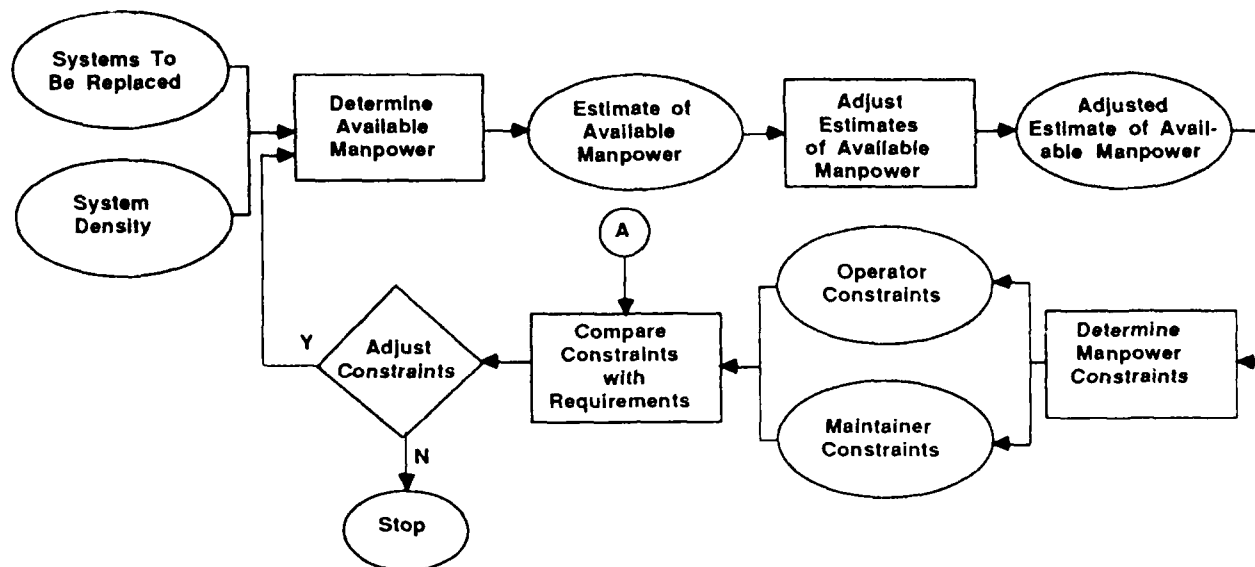


Figure 1. Overview of M-CON Aid Logic

The M-CON Aid determines manpower constraints in a five step process. First, the total pool of available manpower is determined. Initially, this total pool is set equal to the total manpower currently associated with the systems to be replaced. (Later, additional sources of manpower may be identified).

The total manpower for each system to be replaced is a function of the system density (number of systems), the manpower required per system, and an additional factor which makes an adjustment for the discrepancies between manpower requirements and manpower authorizations.

Estimates for maximum operator manpower are relatively straightforward. Crew size requirements from one or more of the systems to be replaced are multiplied by the density of those systems. The actual crew size constraint for the new system depends on the replacement of new systems for baseline systems (i.e., one-for-one, one-for-two, etc.). The user will have the option of looking at as many replacement strategies they wish.

Estimates for maximum maintenance manpower available to support the new system are more difficult to derive. This is because maintenance personnel in a unit often support more than one type of weapon system. In the M-CON Aid the MARC maintenance data base is used to determine the number of annual maintenance manhours required to support a single baseline weapon system. These manhours are converted to productive (i.e., "Hands-on") maintenance manhours at each maintenance level by MOS and paygrade. Those manhours are then translated into manpower requirements using standard Army manpower determination algorithms. Depot level maintenance is not addressed for several reasons: One, there is a lack of reliable data at that level and, two, much of this maintenance has been contracted out to industry.

Second, the user may apply several different adjustment factors to the pool of available manpower. These adjustment factors reflect the fact that there are several different ways of allocating manpower to the system.

To set a manpower constraint, one can use the manpower actually required to operate and maintain the system to be replaced. However, the personnel actually authorized to a system may be different than the number required. In addition, because of the structure of the Army personnel system, the number of personnel who actually end up working on a system in a particular unit may be different than the number authorized. In addition, the actual number in a particular MOS will vary over time depending on the type of people entering the Army and a particular MOSs share of those different types. The M-CON Aid permits the user to adjust the pool of available manpower requirements for each of these factors (authorizations, actual personnel strength, and projected personnel strength). It also permits users to make adjustments for the differences between total manpower requirements based on MARC and actual unit authorizations.

Third, the manpower constraints for the new system are determined. Manpower constraints per system (i.e., maximum crew size and maximum maintenance manhours) are determined by dividing the total available pool by the system density of the new system. Total manpower constraints are initially calculated using the pool of available manpower determined in the first step. Later constraints are calculated using the adjustment factors from the second step.

Fourth, estimated manpower requirements are entered and compared with the manpower constraints identified in Step 2. The estimated requirements may be derived from the fifth MANPRINT methods aid, a HARDMAN analysis, or another analysis.

Fifth, if the estimated requirements exceed the constraints, the system assists the user in assessing several options for changing the constraints. These options include increasing the system types to be replaced, increasing the number of the systems to be replaced by increasing the units where the replacement will occur, or decreasing the system density of the new system. System density can be lowered either by decreasing the number of units that will get the new system or the number of systems assigned to each unit.

P-CON Aid

The P-CON Aid estimates personnel quality constraints. More specifically, the P-CON Aid estimates the future distribution of key personnel characteristics. These distributions describe the numbers and percentage of personnel that will be available at each level of the personnel characteristics. The P-CON Aid also provides guidance to help Army analysts and contractors understand the impacts of setting constraints at different personnel characteristic levels. For example, the P-CON Aid will display the levels of performance that can be expected at each of these levels. The user can use the information on expected performance to set personnel constraint levels for each characteristic.

Assumptions

P-CON is designed to be applied to major systems but the P-CON Aid logic is general enough to be applied to any system. The P-CON Aid estimates constraints based on personnel availability. Another MANPRINT method estimates the level of personnel characteristics that will be required to successfully operate and maintain a particular contractor's design. The P-CON Aid does not attempt to estimate MOS qualification requirements -- rather it attempts to estimate system-specific constraints.

Method Overview

First, the P-CON Aid estimates what the future distribution of the personnel characteristics will be. Then, it uses results from analyses of the Project A data base to show what levels of performance are achievable at different characteristic levels. The user may then use the information on both personnel availability and performance to identify minimum acceptable levels for each personnel characteristic. There is an inverse relationship between personnel availability and expected performance. The lower one sets the personnel characteristic levels the greater the probability the Army has of finding someone available to man the new system, albeit with a lower expected performance level. The P-CON Aid allows the user to conduct tradeoffs of both of these critical variables.

The P-CON Aid has options allowing users to : (1) output information on projected personnel characteristic distributions in a format compatible with the MANPRINT Target Audience Description (TAD), and (2) compare projected distributions for two different MOSs.

Method for Projecting Personnel Characteristic Distributions

Figure 2 provides an overview of the method used to project personnel characteristic distributions. To project personnel characteristic distributions, the P-CON Aid includes an algorithm that projects how future accessions will be distributed among key subpopulations. The key input to this algorithm is data on the projected propensity to enlist for these key subpopulations. Once the P-CON Aid has produced the projected accession distributions, it then applies a modified version of the Army's Manpower Long Range Planning Model (MLRPS) to estimate how the subpopulations will be distributed in the upper paygrades in future years. The MLRPS flows the accession subpopulations through the personnel system using existing transition rates (the user can change these rates to reflect future personnel policy changes).

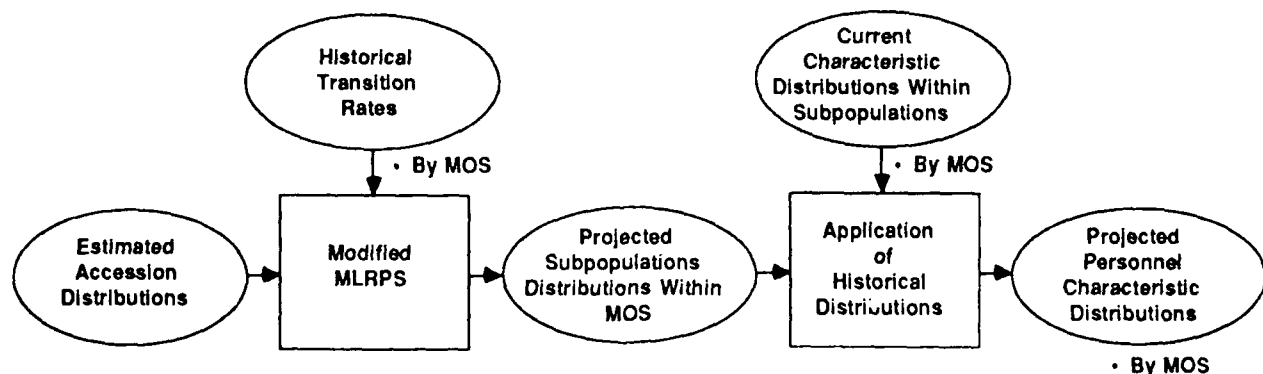


Figure 2. Projecting Personnel Characteristic Distributions

Once the projected distributions of the subpopulations have been estimated, the P-CON Aid uses data on the current distribution of personnel characteristics within these subpopulations to estimate the future distribution of the personnel characteristics levels.

Guidance for Contractors

The P-CON Aid provides guidance to help Army and contractor personnel understand the impact of the personnel characteristic levels on human performance. The P-CON Aid has libraries for 21 different types of systems. One of these libraries describes the generic functional tasks associated with each system type (these generic functional tasks are used in the first MANPRINT methods aid to

describe system requirements). The library also assigns each of the functions to one of the task types in the P-CON Aid task taxonomy. As a result of an analysis of the performance data in ARI's Project A data base, separate regression equations have been developed that predict performance for each task type. In the P-CON Aid, these equations are used to predict what level of performance can be expected at each of the characteristic levels.

To provide additional guidance to contractors, the P-CON Aid provides a set of libraries describing (a) human cognitive and perceptual limitations, and (b) population distributions on key anthropometric and strength variables. All of this information is available in other sources -- the goal in putting it in the P-CON Aid is to put all of the key information needed to describe personnel characteristic constraints in a single easily accessible source.

PER-SEVAL AID

The objective of the PER-SEVAL Aid is to find what level of personnel characteristics is needed to meet system performance requirements given a particular contractor's design, fixed amounts of training, and the specific conditions of performance under which the system tasks will be performed.

The personnel quality requirements produced by the PER-SEVAL Aid will be used to evaluate a contractor's design. These evaluations may be made as early as the proof-of-principle phase of the acquisition process and would probably be continued in subsequent phases. The primary users of the PER-SEVAL Aid would be the Directorate of Combat Personnel who provide input to the Cost and Operational Effectiveness Analysis (COEA), the Logistic Support Analysis (LSA), and the logistic division or group on the program manager's staff who develop manpower and personnel information for the LSA.

The PER-SEVAL Aid has three basic components. First, the PER-SEVAL Aid has a set of performance shaping functions that predict performance as a function of personnel characteristics and training. Second, the PER-SEVAL Aid has a set of stressor degradation algorithms that degrade performance to reflect the presence of critical environmental stressors. Third, the PER-SEVAL Aid has a set of operator and maintainer models that aggregate the performance estimates of individual tasks and produce estimates of system performance.

Method Overview

Figure 3 provides an overview of the PER-SEVAL Aid steps. The user begins an application of the PER-SEVAL Aid by applying the performance shaping functions using the mean level of the personnel characteristics and the estimated amount of training for the new system. Then, these performance estimates are input into the stressor degradation algorithms where performance is degraded to reflect the presence of the stressors. Next, the revised task performance estimates are input into the operator and maintainer modes which aggregate them to produce estimates of system performance. Then, required performance is compared with estimated performance at either the task or system level (the user selects the level). If performance is adequate, the PER-SEVAL Aid stops. Otherwise, the personnel characteristics are incremented or decremented and the entire process is iterated until the required performance levels are met.

The performance shaping functions predict task performance as a function of personnel characteristics and training. Separate functions are provided for different types of tasks. Two types of training variables are used in the performance shaping functions -- frequency and recency of practice. The primary data source for developing the functions were results from a regression analyses from the Project A data base. However, functions for heavy, physical, or gross motor tasks were taken from existing literature. The stressor degradation algorithms lower task performance to reflect the presence of six critical

stressors: heat, humidity, cold, noise, Mission Oriented Protective Posture (MOPP) gear, and continuous operations (lack of sleep). The PER-SEVAL Aid integrates stressors degradation algorithms already available in the human factors literature and organizes these algorithms by the task types in our task taxonomy. The PER-SEVAL Aid uses an algorithm developed by the Army's Ballistic Research Laboratory to aggregate the impacts of multiple stressors.

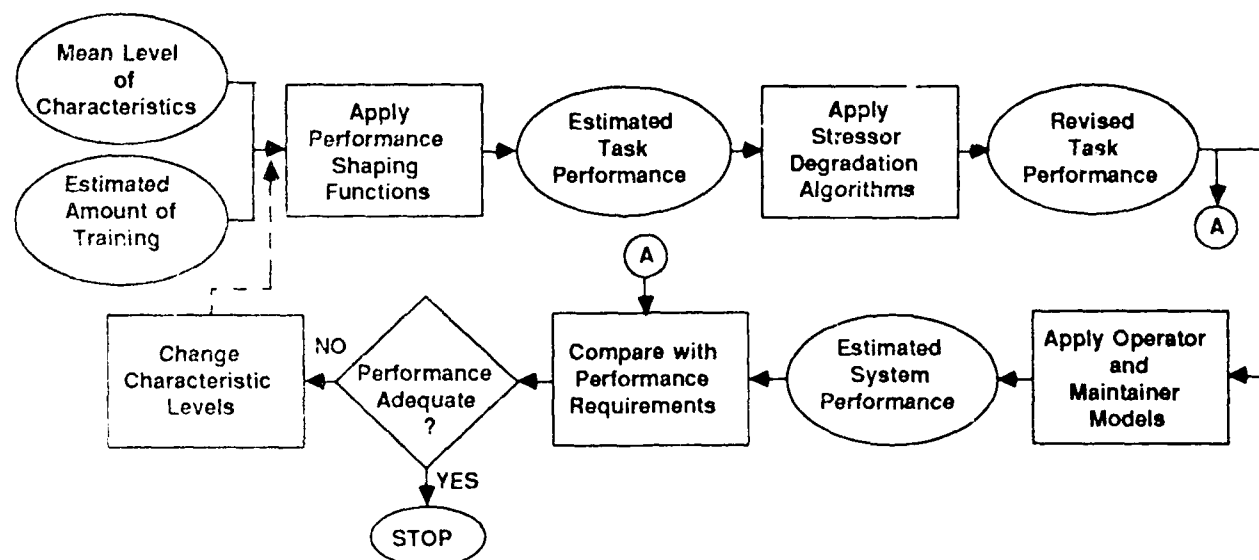


Figure 3. Overview of PER-SEVAL Aid Logic

Task Taxonomy

The PER-SEVAL taxonomy is primarily an expansion of Berliner's (1966) task taxonomy. However, an attempt was also made to incorporate Wicken's (1981) structure for processing resources. These two structures are quite congruent with one another. Task types were eliminated which, while possible to imagine on a theoretical basis, seldom occur in the Army.

Need for Tools to Estimate System Performance

The SPARC Aid will set performance requirements at several levels. At the highest level, it will set performance requirements for each mission associated with the new system. At the lowest level, it will set performance requirements for low-level functional tasks. Operational activities will be broken down to the lowest functional task level while maintenance activities will be described only at a high level.

The PER-SEVAL Aid has separate simulation models for operators and maintainers. The operator model simulates the tasks that operators will perform during the mission. It produces estimates of mission time and probability of mission success. The model calculates total mission time by keeping track of the time spent on individual tasks on a particular mission. The model aggregates task accuracy estimates into function accuracy estimates and combines the function accuracy estimates to produce estimates of probability of mission success.

The maintainer model simulates maintenance task performance. The maintenance model also processes information on component reliability and the time to perform each maintenance task and produces estimates of system reliability, availability, and maintenance manhour requirements.

MANPRINT Evaluation of the Theater Army Medical Management Information System
Norman D. Smith
John R. Tiffany
U.S. Army Research Institute

INTRODUCTION

Background of TMMIS/TMMIS-D

In late 1979, the Office of the Surgeon General (OTSG) recommended development of the Theater Army Medical Management Information System (TMMIS) to meet anticipated information requirements of field medical units. A Mission Element Need Statement (MENS) was approved in January, 1981. In February 1983, a contract was awarded to develop TMMIS to serve the Division Medical Supply Office (DMSO) and higher echelon medical units up to the Medical Group at Theater level. In October 1985, the contract was modified to extend TMMIS to serve the DMSO and lower echelon units down to the medical platoons within combat and combat support battalions. The Division level and below variant of TMMIS was named TMMIS-D.

The TMMIS/TMMIS-D systems were developed to meet the needs of medical commanders by providing timely, accurate and relevant information on the status of patients, medical units and medical supplies on the battlefield. TMMIS is defined as an automated, on-line, interactive, microcomputer system that manages combat medical information. The system is designed for wartime operations, but includes the automation of peacetime functions that can be easily suppressed in transition to war.

Within TMMIS-D are the subsystems Medical Logistics (MEDLOG-D) and Medical Patient Accounting and Reporting (MEDPAR-D). MEDLOG-D accommodates the management of medical supplies, medical assemblages, and bio-medical equipment maintenance for divisional Table of Organization and Equipment field medical units. It is designed to operate at the Medical Platoon and Forward/Main Support Medical Company levels. MEDPAR-D supports medical treatment organization/element commanders and/or special staff in the management and accountability of patients. Individual patient data and medical information is accumulated to monitor the status of troop health and medical resource usages.

TMMIS has similar functions for MEDLOG and MEDPAR but their operational area is from Corps upward. TMMIS has the additional subsystems Medical Blood (MEDBLOOD) and Medical Regulating (MEDREG). MEDBLOOD provides for the management and distribution of blood products within the theater and between the theater and Blood Transshipment Centers and operates at Corps level and at echelons above Corps. MEDREG supports the decision process that matches a patient's medical requirements with available transportation and medical treatment resources.

ARI Test Support

In October 1987, the U.S. Army Research Institute Field Unit, Fort Hood, was requested by the Test Agency of the U.S. Army Communications Electronics Board (USACEBD) to provide MANPRINT support for the TMMIS/TMMIS-D Initial Operational Test and Evaluation (IOT&E) scheduled for January-March 1988. Prior to the USACEBD request, apparently no expert MANPRINT assistance had been requested or provided for the TMMIS/TMMIS-D project. From inception through completion of IOT&E, USACEBD and ARI were the only two agencies not subordinate to the Army Office of the Surgeon General to evaluate the project.

METHOD

Test Personnel Requirements

The IOT&E was conducted at Fort Lewis, Washington. Soldiers from several of its medical units participated. The test was divided into two test periods of 10 days each. The first period tested TAMMIS-D and the second period tested TAMMIS.

The Test Design Plan (TDP) prepared by the U.S. Army Communications and Electronics Board (USACEBD) specified the type of MOS required for each of the subsystems of TAMMIS/TAMMIS-D representing the "slice" of the unit included in the test.

Operator and Data Collector Selection Procedures

Medical units of the 9th Infantry Division and I Corps, Ft. Lewis, Washington provided soldiers for the test. Selection was a joint undertaking between the military medical units and the USACEBD based on MOS requirements of the test, availability of personnel during the test period, and utility of the training to the supplying units.

TAMMIS-D. Twenty soldiers were assigned and trained preceding the Initial Operational Test and Evaluation (IOT&E) which was conducted by the USACEBD test team. On completion of training, six operators and four data collectors were selected for this portion of the IOT&E. In order to have sufficient data collectors for the TAMMIS IOT&E which was to follow, an additional six from the trained group of 20 also received instruction in data collection from the USACEBD test team. The four remaining trained personnel were assigned other duties for administrative reasons.

The TAMMIS-D trainees were evaluated by the instructors from Health Care Systems Support Activity (HCSSA) during the training. The six operators were selected on the basis of this evaluation and their GT scores. The criterion for selection was predicated on the doctrine enunciated by the surgeon general that all medical personnel would be required to operate the TAMMIS/TAMMIS-D system. The USACEBD test team and ARI concluded that for a realistic test of this criterion, operators should be selected from the trainee group who fell at or below the average performance at the completion of training and who had GT scores at or below average (100).

The selection of data collectors was based on the criterion that those who understood the system best would be better able to detect deficiencies in the system. Hence, those trainees rated above average on performance during training on the TAMMIS-D system were selected.

TAMMIS. In like manner to TAMMIS-D, 20 soldiers were trained for this test. The selection by the 9th ID of personnel was heavily influenced by the division's interest in training those personnel who would be the first to use the TAMMIS system upon fielding. The selection method was based on the knowledge that all training, at least initially, would be conducted by the newly automated units training themselves. The consequence to the IOT&E was the elimination of the selection criteria established earlier by the USACEBD, therefore the lower end of the performance curve in training no longer determined who should be operators during the test. The test required twelve operators and eleven data collectors. Only two data collectors were used from this training group to make up the compliment of eleven that were needed. Nine were carried over from the TAMMIS-D test; consequently, these latter nine had not had training on the TAMMIS system.

Test Personnel Representativeness

A comparison of operator GT and CL composite ASVAB scores was made with the mean of these scores for their PUS to estimate the comparability of the test population to the population as a whole. The average GT composite score

for TAMMIS-D operators was 100.5 while the average GT composite score for TAMMIS operators was 112. One TAMMIS-D operator fell one standard deviation below the mean and four TAMMIS operators exceeded one standard deviation above the mean. A similar finding for the CL composite was observed with the TAMMIS-D operators' average CL being lower than that of the TAMMIS operators (98 vs. 106.88). Two TAMMIS-D operators fell more than one standard deviation below the mean.

Data Collection

The TAMMIS-D test required four data collectors, one stationed at each of the positions noted in Table 1.

Table 1

TAMMIS-D Data Collector/Player Organization

Bn Aid Station			
MEDPAR-D		91A Operator	Data Collector
MEDLOG-D		91A Operator	
Bde Clearing Station			
MEDLOG-D		91B Operator	Data Collector
		91A Operator	
MEDPAR-D		71G Operator	Data Collector
Bde Surgeon			
MEDPAR-D	C&C	91B Operator	Data Collector

In order to control for bias and to give all data collectors experience, they were rotated on a daily basis. For the TAMMIS test two additional data collectors were added from the personnel who had received TAMMIS training. These were the only data collectors who had been trained on the TAMMIS/TACCS system. The TAMMIS test required 9 data collectors, one stationed at each of the positions shown in Table 2.

Table 2

TAMMIS Data Collector/Player Organization

Combat Support Hospital			
MEDBLOOD		92B Operator	Data Collector
MEDMAINT		35G Operator	Data Collector
MEDPAR		71G Operator	Data Collector
MEDSUPPLY		76J Operator	Data Collector
		76J Operator	Data Collector
Medical Group			
MEDBLOOD		92B Operator	Data Collector
		71G Operator	Data Collector
MEDREG		71G Operator	Data Collector
MEDPAR	C&C	71G Operator	Data Collector
MEDSUPPLY	C&C	76J Operator	Data Collector
Division Surgeon			
MEDPAR	C&C	91S Operator	Data Collector
Blood Module			

MANPRINT UTILITY AT THE EARLY STAGES OF SYSTEM DESIGN

A test and evaluation yields the most instructive information when it is possible to include all the major MANPRINT issues during the early stages of system design. MANPRINT issues are often difficult to include as development moves toward the IOT&E stage. The result is that issues such as workload cannot be addressed and comparative measures of the new and old systems cannot be taken. This argues strongly for the inclusion of MANPRINT considerations throughout the developmental cycle.

Part I of the following discussion points out some major considerations related to the TAMMIS/TAMMIS-D system that could have been more adequately addressed if MANPRINT concerns had entered the development cycle earlier. In addition, a brief discussion is presented of the data collection procedures that were devised to attempt to offset the short-comings in the design of the test. Part II presents a synopsis of the MANPRINT findings from the IOT&E.

Part I: Test Design

Work Load. One presumption underlying TAMMIS/TAMMIS-D development is that the addition of computer managed information will keep commanders better informed and informed in a more timely manner. To assess this hypothesis, a test design that measures both the accuracy and the speed of the input at the work station is necessary. Of equal or greater importance, however, is the effect of task realignment on the primary task of the combat medic, that is, attending to casualties in wartime. Possible reduction in the time available to perform the primary task is actually the limiting condition to TAMMIS/TAMMIS-D. To assess this condition, the test design just simulate to some degree the actual combat work load. The current IOT&E did not do so. Hence, the major question of whether a medical soldier is capable of entering data into TAMMIS/TAMMIS-D in a timely and accurate manner, while performing his or her primary task, is not answered. Further, no estimate of the impact of 24 hour operations was made.

System Comparability. The test did not take advantage of the opportunity to compare the TAMMIS/TAMMIS-D with the existing system under simulated wartime conditions. The existing system requires medics to fill out forms by hand. The critical echelon for consideration of this question is at the battalion aid station where TAMMIS-D will operate. In addition, at present a back-up of all information is required and that back-up is the hand written form. The soldier must both enter data in TAMMIS-D and keep the forms up to date. It is important to compare the current system with a TAMMIS-D driven system to determine how the soldier functions in each.

Both of the factors discussed would provide the information necessary to estimate whether TAMMIS/TAMMIS-D adds significantly to the task or reduces the medic's work load. If it is an added burden, a dedicated operator is one solution. The IOT&E as designed could not produce data pertinent to these issues.

Performance Measures Used in IOT&E. The two most frequently used measures by the USACEBD were time and errors. Start and completion times were recorded by data collectors of all events. However, as a result of the fact that requirements contained in the Independent Evaluation Plan, Operational Mode Summary/Mission Profile, and the test scenario were not in agreement with each other, the data collection forms were organized in such a manner that discrete report production times could not be extracted. The result of the mismatch of program documents and data collection instruments was the inability to be able to examine the relationship between personnel performance and aptitudes. The error measure was based on the number of incorrect characters appearing in selected fields on predetermined critical reports. As a measure of the TAMMIS/TAMMIS-D system, it showed that the software and

hardware accepted the input. However, far less elaborate and costly assessments of this component of the test could have been devised. The measure also served as a test of operator entry aptitudes composed of manual dexterity, vision, and previous typing experience or training, all of which could also have been assessed by other less costly means.

Assessment of the Repeated Interview Method. Interviews were conducted after each of the test days with all participants. This procedure was used to estimate the amount of learning that takes place after testing begins and to identify system problems that are not resolved during the dynamics of the test situation. This analysis was based on data found in Table 2.

Comments obtained during the first three days of the test were collated within the MANPRINT areas of manpower, personnel, training, human factors, health and safety and presented at the end of the test. An arbitrary acceptance level of 50% or greater was established for determining whether a problem had persisted throughout the test based on the combined responses of operators and data collectors. The following results were obtained.

Table 3

Comparison of MANPRINT Problems Identified Days 1-3 with Last Test Day

Manprint Areas	Number of Problems Identified	
	1st 3 Days	Last Day (50% Criterion)
Manpower	2	0
Personnel	4	4
Training	10	9
Human Factors		
Software	20	6
Controls and Equipment	18	8
Manuals	8	6
Health	1	1
Stress	3	3
Safety	3	1

Issues pertaining to personnel (aptitudes and attitudes) and training were identified early in the test and remained essentially unchanged until the end. However, issues of human factors having to do with the software, equipment, and manuals were greatly reduced by the test's end. These findings suggest that soldiers know very early in the test what skills are needed and whether their training is adequate. The findings also showed that learning continued and many of the human factors issues disappeared after a few days of practice with the system software and manuals.

Part II: Synopsis of Test Findings

All operators complained that the data processing rate was extremely slow and suggested that the mobile medical units would find the TACCS to be unusable. The slow processing time and time taken to print out lengthy reports would probably not be accommodated by a unit that has to periodically pack up and move. Further, the bulk of the equipment meant to operators that something else in the limited operations space would have to be sacrificed. If a hand written method was required in any case, the TAMMIS hardware would probably stand little chance of being used.

(1) The training packages used to train operator personnel for both TAMMIS and TAMMIS-b prior to the start of the tests were inadequate; and provisions must be made for retention or sustainment training.

(2) All operators should receive an introductory course in computer operation and all should have some ability in typing.

(3) Workload considerations were inconclusive due to confounding conditions in the experimental design.

(4) Manuals must be revised so that persons who are not computer literate can readily locate information and understand the instructions for taking corrective action.

(5) Fourteen specific problems with software must be corrected.

(6) A "Help" line must be established with system experts who have both expert knowledge of the system and ability to communicate effectively with the target audience.

(7) No safety hazards were identified.

(8) The TACCS hardware should be replaced or upgraded to correct:

- (a) inordinately slow operating time,
- (b) length of time to boot up the system,
- (c) dim CRT lighting or contrast, and
- (d) bulk.

Testing Aquila, The Army's Remotely Piloted Vehicle: Lessons Learned

Edwin R. Smootz and Nigel R. Nicholson
Army Research Institute
Fort Hood Field Unit

Introduction. Aquila, the Army's Remotely Piloted Vehicle (RPV), is a relatively small unmanned aircraft that is remotely controlled from a ground station and is designed to carry a payload that can serve target acquisition, designation, reconnaissance and surveillance functions. Development of the system can be traced back to 1971 when the Defense Science Board recommended that the Army establish a program which would apply mini-RPV technology to the fire support functions of target designation and artillery adjustment. The concept of using pilotless aerial vehicles for military purposes was investigated as early as World War I, albeit with little success (Joint Electronic Warfare Center, 1986). It was looked at again in World War II, and has been under some degree of research and development ever since. Much of the early success was with fairly large target drones which simulated flying aircraft and were used for gunnery exercises. Eventually, the idea of modifying such drones by hanging cameras on them emerged, and this combined with evolving technology created many possibilities for military application. For example, the development of the electronic digital computer, progressive miniaturization of components, and other related developments in the 1960's and 1970's, provided the possibilities for building relatively small air vehicles which could continuously send real time digitized data to a ground station. In addition, it allowed controllers to actively control the flight path or permit the air vehicle to follow a preprogrammed flight path of relatively long duration. Finally, such small vehicles could be easily launched and recovered, and for some requirements procured cheaply enough to be expendable.

It was in this technological milieu, combined with an increasing concentration of Soviet ground-based air defense weapons posing a high threat to piloted reconnaissance aircraft, that the Defense Science Board made its recommendation in 1971 to develop mini-RPV's for reconnaissance and target acquisition. In September, 1974, the program formally got underway when the U.S. Army Training and Doctrine Command (TRADOC) and the U.S. Army Materiel Command (AMC) signed a letter of agreement to develop an RPV system that could demonstrate how RPV's could be used by ground commanders in reconnaissance, target acquisition, laser designation of targets, and artillery adjustment. This effort evolved into the development of a prototype system which came to be known as Aquila and underwent three series of tests: Developmental Test (DT) II in 1985-1986, (Cozby, 1986), Operational Test (OT) II in 1986-1987 [U.S. Army Operational Test and Evaluation Agency (OTEA), 1987], and Force Development Test and Experimentation (FDTE) in late 1987 [U.S. Army TRADOC Test and Experimentation Command (TEXCOM), 1988]. Following the conclusion of the FDTE, the Army decided to halt further development and cancelled acquisition plans for Aquila. The Army's RPV efforts have since been combined with those of the other military services under a Joint Program Office (JPO) for Unmanned Aerial Vehicles (UAV), and the few Aquila systems that were initially built for test and evaluation purposes are currently serving as a test bed for the joint program.

The extensive testing and evaluation that Aquila underwent during its multiyear development has provided us a great deal of useful information about how unmanned aerial systems should be designed, operated and employed. The

remainder of this paper will focus on some of the problems and issues in the human factors area that were brought to light during this period. Specifically, we will discuss three topics: (1) Target Acquisition, (2) Mission Planning, and (3) Manpower and Personnel Selection.

Target Acquisition. One of the major findings of OT II was that Aquila crews too often failed to detect targets. Examination of the results from OT II (OTEA, 1987) indicates that overall detection rate was around 17% (24% of moving targets and 13% of stationary targets). Several uncontrolled extraneous factors contributed to this low detection rate. In addition, Aquila crews were required to search too large an area during the test. The mean area reconnoitered during OT II was 17.6 km^2 , which exceeded the 16 km^2 area the Army considered to be the maximum for a typical three hour Aquila sortie (U.S. Army Field Artillery School, undated). It also became obvious during the test that searching for targets from an altitude of 1500 meters with a 20° field of view is very much like looking at the world through a straw, and is a very difficult and tedious process.

Between the completion of OT II and the beginning of the FDTE, the Army reevaluated Aquila search requirements using data from OT II and experience learned from other systems and concluded that it was not feasible to search so large an area for targets in general. Instead, it seemed better to use other intelligence sources to indicate approximately where targets may be, and to use such information to cue Aquila crews to search a more restricted area.

This change was incorporated into the operating procedures established for the FDTE. In addition, software was developed to assist the operator in performing reconnaissance (TEXCOM, 1988). A bookkeeping function was developed which automatically kept track of the amount of an area that had been searched. During OT II operators easily lost track of where they were when they took manual control of the system, thus making it difficult to systematically search an area. With the new bookkeeping function an operator could take manual control of the system, leave the area to perform an emergency search some place else, and, providing he did not enter a new waypoint, come back and automatically pick up the search where he had previously left off.

The other software improvement was an automatic search routine known as the step-stare technique. This technique focused on a given portion of the search area for a fixed amount of time, depending on the amount of clutter in the area, then automatically moved to and focused on an adjacent area with about a 10% overlap, and thus, provided for the systematic search of a whole area. This change increased the time required to search an area, however, from 12 to 35 minutes per km^2 . Consequently, the Army reduced its maximum area search requirement to 6 km^2 per sortie (U.S. Army Field Artillery School, 1987).

All of the above changes, i.e., reduction in the maximum size of an area to be searched, cueing of the general target area, incorporation of an automatic bookkeeping function, and utilization of the automatic step-stare search process, were incorporated into the Aquila system prior to the FDTE with the intention of increasing the target detection rate. The results of the FDTE indicated that these changes did help. Ninety-eight percent of all moving target arrays and ninety-four percent of all stationary arrays that were not camouflaged were acquired by the Aquila operators during the FDTE (TEXCOM,

1980). However, no target arrays that were camouflaged were acquired. Thus, the results were very clear-cut, showing a definite ability to detect moving and stationary, uncamouflaged targets, and a definite inability to detect camouflaged targets that were camouflaged with either artificial or natural material.

There were additional target acquisition problems which had been identified during the OT II, and still posed a problem during the FDTE. For example, operators found it very difficult to detect targets that were in shadows, as often happens when one camouflages a vehicle by parking it under or near a tree. Dr. Aaron Hyman (1987), recently of the Army Research Institute, identified a potential contributor to this problem, along with a possible solution. Although the camera in Aquila was of very high quality, the circuitry associated with the automatic gain control and the automatic level control was designed to clip the upper and lower 10% of the illuminance falling on the photocathode. A linear transfer function described the relationship between the middle 80% of the photocathode illumination and the output signal that was transmitted to the ground control station and translated into various shades of grey on the video display. When an Aquila operator was searching a scene, the camera typically received a wide range of input luminances, the lower of which often came from dark or shadowy areas. But since the lower 10% of the input luminances was clipped, the information about objects in the shadows was not transmitted to the video monitors on the ground.

Dr. Hyman suggested a number of ways to solve this problem. A procedural solution would be to use a narrow field of view (FOV), such as 2.7° , to search shadowy areas so that the dynamic range of input luminance spans just the shadowy area. However, this is only feasible for short periods of time. Of the three Aquila FOV options, 2.7° , 7.2° , and 20° , operators tended to use the 7.2° and 20° FOV's much more than the 2.7° FOV because of the tremendous decrease in the amount of terrain they could observe through a narrow 2.7° FOV, and the corresponding increase in the amount of time needed to search an area. An engineering solution to this problem would be to provide the operator with the option of switching, as circumstances dictate, to circuitry which would not clip the lower input luminances and would employ a non-linear transfer characteristic with a high gain at the lower end of the input luminance distribution. This could be used to, in essence, amplify and detect any signals that existed in shadows.

It might be noted that Forward Looking Infrared (FLIR) cameras are seen by many observers as a solution to this problem of detection of targets in shadows. However, it should be noted that FLIR, while having many advantages, does not always produce as clear a picture as one would like. When contrast between two objects is low, it is still difficult to distinguish them, and the Army currently plans to provide for both daylight TV and FLIR as potential payloads on its UAV system that is in the initial stages of development. Thus, the problem of detecting targets in shadows with daylight TV cameras is one which must be addressed.

Another aspect of the problem of detecting camouflaged targets concerns the combination of slant range and FOV for searching an area. Dr. Hyman's communication with various experts in the field of imagery analysis indicated that the acquisition of targets in clutter, such as vegetation, is a somewhat different process than the acquisition of targets in open terrain. In the

latter situation an observer typically detects something, and then after obtaining better resolution by moving closer to it or magnifying it, recognizes it as a particular type of object. In clutter, however, it appears that detection of an object or target does not occur until resolution is sufficient to also recognize the target. Previous research (Johnson, 1958) has shown that recognition of military vehicles with a probability of 0.5 requires about eight TV lines; thus, in clutter, it appears that about eight TV lines are needed to also detect a target. Dr. Hyman calculated that with Aquila's 7.2° FOV, a slant range of 1.66 km was required in order to obtain enemy target representation (e.g., a tank) of eight TV lines. However, Aquila doctrine suggested a 2.5 km slant range for area search with the 7.2° FOV. Thus, it is possible that some targets in clutter were not detected during testing of Aquila because adequate target resolution was not obtained at the video monitor due to the non-optimal combination of slant range and FOV.

One way to resolve this problem would be to use a narrower FOV. This would give more TV lines to a target of given size. Another alternative would be to simply use a shorter slant range, such as the 1.66 km slant range mentioned above, although this would require flying below that altitude which had been determined as the minimum for sustained survivability.

Mission Planning. One problem uncovered during the OT II was that Aquila crews accepted from higher headquarters mission requests that exceeded the capabilities of the system (Test and Experimentation Command, 1988). This typically took the form of crews accepting a mission to search an area larger than Aquila could reasonably handle during a three hour sortie. The exact cause for this was not empirically determined, but it was thought to be due to several factors, to include lack of understanding of Aquila's limitations by the crews, and a tendency to be unquestioning when receiving mission orders from a higher headquarters. As a result, the training program was modified prior to the FDT&E to give Aquila crews training on system limitations and the need to resolve conflicts with higher headquarters over nonexecutable mission requests. During the FDT&E crews were given ten incomplete mission orders and nine nonexecutable mission orders. They were scored on whether they resolved those conflicts. Scoring occurred in two ways. First, a subject matter expert (a senior warrant officer) observed the crews during the three phases of a mission (i.e., receiving the mission order at higher headquarters, planning the mission in the Ground Control Station, and executing the mission by flying the air vehicle and searching for targets) and judged whether or not they resolved mission request conflicts with higher headquarters (TEXCOM, 1988). Second, videotapes of crew actions during all three phases were scored by independent evaluators to determine the frequency and distribution of crew actions involving: a) requests for clarification of missions from higher headquarters, and b) explaining system capabilities to higher headquarters (Nicholson, Deignan, and Smootz, 1988).

The results were conclusive. The subject matter expert determined that crews identified and adequately resolved all 19 conflicts presented to them. The videotape analysis showed that this conflict resolution behavior was about evenly spread out over all three phases of mission execution. However, as a proportion of the communications activity occurring during each phase, it was found that it comprised most of the communication (88%) between Aquila crewmen and higher headquarters while receiving the mission, but only 8% of such communication during mission planning, and less than 1% during actual flight.

The lesson to be learned here, however, is that specific training in system limitations is extremely important, especially given that searching an area with UAV's has proven to be a very slow and tedious process. Individuals in higher headquarters who are tasking UAV's may easily overestimate their search capability and it is crucial for UAV crews to be trained to expect this and be prepared to resolve such problems with higher headquarters.

Manpower and Personnel Selection. The third topic we wish to discuss focuses on the manpower and personnel requirements for Aquila. There are two basic issues here: the number of personnel required for manning the system, and the aptitudes required for performing the various tasks associated with operating the system.

With respect to the number of personnel required, information from OT II and an independent analysis of manpower, personnel and training issues (ARI, 1987) indicate that Aquila was far more manpower intensive than originally planned. Maintenance manpower turned out to be a particular problem and will be addressed by Dr. Stewart in the following paper. Operations manpower had its own set of problems also. For example, the OT II test criterion for preparing and launching an air vehicle after receiving a mission order was a maximum of 60 minutes on 80% of the trials. However, the time criterion was met on only 44% of the trials. Part of the failure to reach criterion can be attributed to the high number of launches that were aborted because of maintenance problems indicated by the built-in-tests. In fact, throughout OT II an average of 2.2 launch attempts were made for every successful launch. Nevertheless, the eight man launch and recovery section was usually able to adequately handle the launch when equipment did not malfunction, and in many cases got a launch off in ten minutes. However, the OT II only tested operations during daylight hours. As mentioned earlier, the Army plans to include a Forward Looking Infrared (FLIR) payload on future UAV's and thus provide a 24-hour continuous operations capability. Given the high degree of equipment malfunction that kept crews busy, and the requirement to operate around-the-clock, one is logically led to ask whether an eight-man crew could continue to adequately function for very long. Perhaps two four-man crews could, by working 12-hour shifts, but then one must consider that many other duties arise in combat, such as perimeter security. Unfortunately, no data were collected on this question during any of the tests conducted, but it is one which is obviously very important and we plan to examine it in evaluating manpower requirements for future UAV systems.

A related question that emerged, (and also one on which data are scarce), concerned the skills and aptitudes required for operating Aquila. A basic problem existed in that the skills required to operate a system like Aquila were distinct enough to require a unique military operational specialty (MOS). However, since the system was not a high volume system (the Army only planned to acquire nine Aquila systems requiring about a thousand troops), the density of the MOS was low. This situation was complicated by the fact that soldiers with the Aquila MOS were assigned to one of two distinct jobs: either operating the launch and recovery system, where requirements were rather physical, or operating the control station, where requirements were more cognitive and perceptual. It could be argued that two MOS's should have been created for these jobs, but doing so would have meant that each MOS would contain only 400-600 troops, a very low density indeed. The Army personnel system has difficulty managing low density MOS's. The situation was compounded

by the fact that not much empirical data exists on the types of aptitudes actually required for operating the control station, so it was difficult to make a firm decision as to whether or not that job should require a distinct MOS. This is a fundamental issue we plan to address in future UAV efforts.

In summary, the Army's experience in testing and evaluating Aquila revealed a number of human performance problems in such areas as target detection, mission planning, and manpower and personnel selection. These are areas which will continue to be investigated in future Research, Development, Test, and Evaluation (RDTE) efforts related to UAV's.

References

- Cozby, R. S. (1986). Test report: Development Test II of the Army Aquila Remotely Piloted Vehicle (RPV) system (U) (Pub. No. USAEPG-FR-1253). Fort Huachuca, AZ: U.S. Army Electronic Proving Ground. SECRET.
- Hyman, A. (1987). Initial examination of the imaging system/operator interface for the Aquila. Memorandum for Director, Systems Research Laboratory, U.S. Army Research Institute, Alexandria, VA.
- Johnson, J. (1958). Analysis of image forming systems. Image Intensifier Symposium. U.S. Army Engineer Research and Development Laboratories. Fort Belvoir, VA.
- Joint Electronic Warfare Center. (1986). Unmanned vehicle systems remoting technology study (U) (JDN 1-86). San Antonio, TX: Systems Engineering Directorate. SECRET.
- Nicholson, N., Deignan, G., and Smootz, E. (1988). Remotely Piloted Vehicle (Aquila): Force Development Test and Evaluation: Army Research Institute Fort Hood Field Unit Evaluation (WP FH 8804). U.S. Army Research Institute Field Unit, Fort Hood, TX.
- U.S. Army Field Artillery School. (undated). Remotely Piloted Vehicle (Aquila) operation (Draft Field Circular 6-RPV). Fort Sill, OK.
- U.S. Army Field Artillery School. (1987). Remotely Piloted Vehicle (Aquila) operation (Draft Field Circular 6-RPV). Fort Sill, OK.
- U.S. Army Operational Test and Evaluation Agency (OTEA). (1987). Independent evaluation of the Remotely Piloted Vehicle (U) (IER-OT-604). Falls Church, VA. SECRET.
- U.S. Army Research Institute. (1987). Evaluation of Manpower, Personnel and Training sections of the Human Factors Engineering Analysis (HFEA) of the Remotely Piloted Vehicle (RPV). Alexandria, VA.
- U.S. Army TRADOC Test and Experimentation Command (TEXCOM). (1988). Remotely Piloted Vehicle: Force Development Test and Experimentation (TEXCOM Test Report FD 0186). Fort Hood, TX.

ESTIMATING MAINTENANCE MANPOWER REQUIREMENTS FOR THE AQUILA REMOTELY PILOTED VEHICLE

John E. Stewart, II
Army Research Institute for the Behavioral and Social Sciences

INTRODUCTION

Overview

HARDMAN and HARDMAN II (Hardware vs. Manpower) are methodologies for estimating manpower, personnel and training (MPT) requirements for emerging systems. HARDMAN was performed on the Lockheed Aquila Remotely Piloted Vehicle (RPV) by the Field Artillery School TRADOC Systems Manager (TSM) under contract with Dynamics Research Corporation in 1983. The original analysis assumed that the RPV battery would be fielded in five autonomous sections.

The 1985 revision of the Target Acquisition, Designation and Aerial Reconnaissance System (TADARS) RPV Operational and Organizational (O&O) Plan imparted more centralization and control to the Aquila battery. The original five sections were superseded by three Forward Control Sections (FCSs) and two rear area Centralized Launch and Recovery Sections (CLRSs), each with its own Ground Control Station (GCS). Each CLRS has a Launch Subsystem (LS), Recovery Subsystem (RS) Air Vehicle Handler (AVH) and five Air Vehicles (AVs). It is the responsibility of the two CLRSs to conduct launch and recovery operations for the entire battery. Only one CLRS, the Primary CLRS (or CLRS 1) has a maintenance shelter (MS) along with three spare AVs. The MS and its crew of four must provide organizational level maintenance for all 13 AVs. Originally, it was intended that each CLRS have an MS.

A reanalysis of the original HARDMAN was required to incorporate these changes. One principal finding of the revised HARDMAN (1985) was that, because of the shift from a 12 to a 24 hour operational scenario, a total of six military occupational specialty (MOS) 13T P9 maintainers was required at the MS instead of the four required by the 1985 TADARS O&O Plan.

However, lower than expected automatic fault isolation (FI) rates obtained by the automated test equipment (ATE) during Development Testing II (DT II) in 1986 and Operational Testing II (OT II) in 1987 indicated that maintenance manpower requirements should be readdressed. Also, it appeared that single point estimates of ATE performance were not as useful as a range of estimates based on expected levels of ATE performance.

Stewart and Shvern (1988) applied HARDMAN II sensitivity analysis to two components of the Forward Area Air Defense

system. Maintenance manpower estimates as a function of automatic fault isolation performance caused the Army to reexamine its original maintenance manpower requirements. The following sensitivity analysis was performed in order to provide the Army with updated information about maintenance requirements for the Aquila RPV, and to illustrate further its efficacy as an adjunct to HARDMAN II.

METHOD

Documentation

The principal data sources were the 1983 and 1985 HARDMAN analyses. The 1985 TADARS O&O Plan (with changes) provided information on projected repair times, usage rates, and operational scenarios. Test data were available from DT II and OT II which provided information on maintenance ratios, the number of repair actions at organizational and intermediate levels, and operational availability estimates from these tests.

The Required Operational Capabilities (ROC) document called for a successful fault isolation (FI) rate of 90%. ATE performance to date has been poorer than this (see Nauta, 1985). During Development Test II (DT II; see Cozby, 1986), the Aquila ATE system only isolated 35% of all faults. During Operational Test II, (OT II; Operational Test & Evaluation Agency, 1987) this rate was slightly less than 20%.

Analytical Approach

The methodology employed in the present analysis was a "top-down" approach which relied on HARDMAN results as a baseline. Because detailed raw data from the HARDMAN were not available, it was necessary to rely on data from DT II and OT II to obtain estimates of AV down time, repair times and maintenance ratios. The wartime operational scenario from the 1985 TADARS O&O Plan allowed for extrapolation to the total RPV battery. The resultant annual maintenance manhours (AMMH) obtained through the top-down approach (7722) agreed closely with those from the HARDMAN (7961 adjusting for the 24 hour scenario).

Repair times. Because of the lack of a similar fielded predecessor, assumptions from various secondary sources had to be used. Mean Time to Repair (MTTR) times using ATE from the revised O&O Plan are wrench-turning times only (see Table 1).

The ATE system is mounted inside the MS. For diagnoses of faults to be carried out, the AV must be partially disassembled, defueled, and then moved inside the shelter, accounting for the day-night differences in MTTR.

Table 1

Projected Mean Times to Repair at Organizational Level

<u>Method</u>	<u>Time of Repairs</u>	
	Day	Night
ATE	30 min	45 min
Manual	90 min	135 min

Operational scenarios. Total AV operating hours for the 24 hour-25 mission scenario would be 54 hours, based on the 1985 TADARS O&O Plan.

Maintenance ratios. From operational requirements and OT II test results it can be inferred that 1.27 maintenance actions per day will be required per air vehicle.

Effects of travel. For the remote CLRS, there should be 6.35 maintenance actions per day anticipated. It is assumed that two maintainers from the maintenance shelter will retrieve the AV.

RESULTS AND DISCUSSION

The sensitivity estimates showing annual maintenance manhours (AMMH), final maintenance ratio (FMR), operational availability (A_0), AVs and maintenance man years (MMY) as joint functions of (a) ATE fault isolation (FI) rates of 90% (similar to the ROC requirement), 40% (slightly better than at DT II) and 20% (similar to OT II), (b) percentage of repairs performed during daytime hours (it was assumed that it would be 50% for the 24 hour scenario; 80% was considered optimal but may not be practicable under realistic conditions) and (c) distance between CLRS 1 and 2, are presented in Table 2. (Distance between CLRS=0 is equivalent to there being two MSs).

Table 2

Aquila RPV Maintenance Sensitivity Analysis

<u>Distance Between Primary and Secondary CLRS (Km)</u>							
	0	4	6	8	10	12	14
<u>ATE FI = 90%</u>							
	(50% daytime repairs)						
AMMH	7961	8579	8888	9197	9506	9815	10124
FMR	.40	.44	.45	.47	.48	.50	.51
A _O	.60	.56	.54	.53	.52	.50	.49
AVs	7.75	7.34	7.13	6.93	6.72	6.52	6.32
MMY	3.32	3.57	3.70	3.83	3.96	4.08	4.22
	(80% daytime repairs)						
AMMH	5998	6464	6697	6938	7188	7447	7715
FMR	.30	.33	.34	.35	.36	.38	.39
A _O	.70	.66	.66	.65	.64	.62	.61
AVs	9.10	8.58	8.58	8.45	8.28	8.06	7.93
MMY	2.50	2.69	2.69	2.89	2.98	3.10	3.21
<u>ATE FI = 40%</u>							
	(50% daytime repairs)						
AMMH	13988	14605	14915	15224	15533	15824	16151
FMR	.71	.74	.76	.77	.79	.80	.82
A _O	.29	.26	.24	.23	.22	.20	.18
AVs	3.77	3.38	3.12	2.99	2.86	2.60	2.34
MMY	5.82	6.08	6.21	6.34	6.47	6.60	6.72
	(80% daytime repairs)						
AMMH	10966	11818	12243	12684	13141	13614	14104
FMR	.56	.60	.62	.64	.67	.69	.72
A _O	.44	.40	.38	.36	.33	.31	.28
AVs	5.72	5.20	4.94	4.68	4.29	4.03	3.64
MMY	4.57	4.92	5.10	5.29	5.48	5.67	5.88

Table 2 (Continued)

Aquila RPV Maintenance Sensitivity Analysis

<u>Distance Between Primary and Secondary CLRS (Km)</u>								
	0	4	6	8	10	12	14	
<u>ATE FI = 20%</u>								
	(50% daytime repairs)							
AMMH	16398	17016	17325	17634	17943	18252	18561	
FMR	.83	.86	.88	.90	.91	.93	.94	
A _O	.17	.14	.12	.10	.09	.07	.06	
AVs	2.18	1.82	1.56	1.30	1.17	.91	.78	
MMY	6.83	7.09	7.22	7.35	7.48	7.61	7.73	
	(80% daytime repairs)							
AMMH	12963	13969	14993	15201	15534	16093	16672	
FMR	.66	.71	.77	.76	.79	.82	.85	
A _O	.34	.29	.24	.23	.21	.18	.15	
AVs	4.42	3.77	3.38	2.99	2.73	2.34	1.95	
MMY	5.40	5.82	6.25	6.33	6.47	6.71	6.95	

CONCLUSIONS

Maintenance support of the Aquila would have posed no problem if the ATE system were to have performed as specified in the O&O Plan. The sensitivity analyses showed that for ATE FI rates of 90%, the maximum mission requirement (of five AVs airborne at any one time) could have been met with the maintenance manpower resources available. Neither repair scheduling, the number of MSs, nor distance between CLRS would have posed a threat to operational capability.

Were FI rates to fall to 40%, the five-AV requirement could still be met under optimal conditions (80% daytime repairs and two MSs); with only one MS, the CLRSs must be no farther than 4 km. apart in order to meet the requirement.

If FI rates approximated OT II results (20%), the mission

requirement could not be met, regardless of assets or scheduling of repair times.

It should be noted that estimates for the present project agree closely with findings of the draft Human Factors Engineering Analysis (Human Engineering Laboratory, 1987) which found that during OT II the MS crew of a battery "minus" had difficulty keeping up with the workload for a single CLRS. The report also expressed doubt that the MS crew would be able to support an entire Aquila battery. If operational requirements of the Aquila are to be met at all, an FI rate of at least 40%, which is slightly higher than that attained at DT II, must be achieved, along with the aquisition of an additional MS for each battery.

The present analysis further underscores the usefulness of sensitivity analysis, which enhances the effectiveness of HARDMAN II. The uncertain performance of ATE and other electronic fault diagnostics make single point estimates, based on optimistic criteria, impractical.

REFERENCES

- Cozby, R. (1986) Development Test II of the Army Aquila Remotely Piloted Vehicle System (U). Ft Huachuca, AZ; U.S. Army Electronic Proving Ground. SECRET.
- Dynamics Research Corporation (1983). Application of the HARDMAN methodology to the Army Remotely Piloted Vehicle (RPV). Pasadena, CA: Jet Propulsion Laboratories Contract No. 956-320.
- Dynamics Research Corporation (1985). A reexamination of support requirements of the Remotely Piloted Vehicle (RPV) system. Wilmington, MA: DRC Technical Report E-10053U.
- Human Engineering Laboratory (1987). Human factors engineering analysis for the remotely piloted vehicle (Draft Report). Aberdeen Proving Ground, MD: U.S. Army Laboratory Command.
- Nauta, F. (1985). Alleviating fleet maintenance problems through maintenance training and aiding research. (NTEC Technical Report MDA 903-81-C-0166-1, AD A 155919.
- Operational Test and Evaluation Agency (1987). Independent evaluation of the Remotely Piloted Vehicle (U). IER-OT-504. (SECRET; NOFORN).
- Stewart, J.E. and Shvern, U. (1988). Application of HARDMAN II methodology to the Army's Forward Area Air Defense (FAAD) system. Proceedings of the Human Factors Society 32nd Annual Meeting, 1117-1121.

NDI: A REAL SHORTCUT TO SYSTEM TESTING ?

John L. Miles, Jr.
Jonathan D. Kaplan

U.S. Army Research Institute
for the Behavioral and Social Sciences

Introduction

Everybody wants "a better way" to do things. Even toward the end of the twentieth century, when it's hard to be a hero (Roche, 1987), the people who win acclaim are those who find better ways: increased efficiency, lower costs, greater productivity, more satisfied clients. Equipping the units of a modern Army is one of the most complex and expensive of processes in America, and that process has attracted a large number of innovators. Both government and industry organizations routinely offer awards and recognition to employees who think up "improvements" to any part of the process. One such improvement is alleged to be the Army's Nondevelopmental Item (NDI) acquisition policy, explained well by Morgan and Klein (1986).

The "better way" of NDI is argued on two compelling grounds: cost and schedule. If a product already has been successfully developed for the commercial market, why should the Army insist on redeveloping the same product with military specifications and standards, frequent design reviews and obviously greater costs? Those arguments are, however, far more compelling when the product is simple (like a flashlight battery) than when it is a complex weapon system (like an air defense missile). Moreover, Hightower (1988) argues that DoD ought to pursue "full mil spec" NDI procurements whenever possible (apparently on the grounds that manufacturers can be convinced to volunteer to use military specifications in the design of products which may have both a commercial and military application). Whether or not that argument is successful, it is clear that all the armed services will be using NDI-type procurement for as many items as possible.

Do "NDI" and "OT" Go Together?

For decades, the U.S. Army has had two general categories or types of system testing: the first focused on the hardware/software, and the second on the "manned system" in a simulated operational scenario. Over the years the names of these two categories have changed ("engineering" testing became "developmental" and then "technical" testing; "service" testing became "operational" and then "user" testing). Name changes were used to indicate either an innovation in policy or some shift in emphasis. But under whatever names, the Army still had to learn whether the equipment itself worked and whether trained soldiers could use it effectively to accomplish their mission. In none of the previous permutations of policy and internal Army organization did anyone argue that either "less testing" or "no testing" would be helpful.

Seemingly one of the starkest innovations of NDI acquisition is that test and evaluation (T&E) is diminished in importance. Recent Army doctrine says:

"An important advantage of NDI alternatives is reduced acquisition time. This is accomplished, in part, by minimizing Army testing on NDI."

and

"No [technical testing] will be conducted unless AMC identifies specific information needs that cannot be satisfied by contractor or other test data sources."

(AMC/TRADOC, pp. 17.36-17.37)

That policy is not in itself alarming--at least until one reads how it is interpreted in an Army project manager's office (where, remember, any rewards to the participants are most likely to center around the management parameters of cost and schedule--not the ultimate field performance of the manned system). In a surprisingly frank article, a deputy project manager commends efforts to "fight off the weenies" (Lehnes, 1987, p. 4), a clearly disparaging term for employees outside the project management office whose responsibilities include efforts to confirm the predicted field performance of the product. His rationale is one familiar to anyone employed in the T&E community for more than a few years: never mind how well it works; let's just buy it as it is, put it in the hands of troops, and worry about "improving" it later (paraphrased from HQDA, 1986, pp. 3-4). However, former Secretary of Defense Caspar Weinberger expressed a substantially different point of view:

...[O]ur defense is underwritten by the quality of the weapon systems [italics his] we provide to our combat forces. We rely on our ability to equip America's defense forces with the high-quality military systems needed to deter a numerically superior adversary... We have adopted many initiatives...among them...a focus on quality that goes beyond performance to include the whole range of factors that determine the value of our hardware to the soldier, sailor and airman in the field...

(Weinberger, pp. 2-3)

The question for those of us who work in the T&E community becomes, "How do we meet those objectives for new weapon systems while working within the Army policy guidelines concerning NDI acquisition?"

A T&E Concept for NDI

The decision to proceed with an NDI acquisition implies that an evaluation of data about the product has confirmed that it is likely to function satisfactorily in the hands of troops in the field. But how does an evaluation reach such a determination when Army testing of the product is officially discouraged? We believe that a combination of one old method and one new method can provide a conceptual framework for effective test and evaluation in NDI procurement.

The old method was introduced as "SIDTC", an acronym for Single, Integrated Development Test Cycle, originally promulgated by a Headquarters AMC letter in 1974. SIDTC, in its simplest form, was a marriage of two ideas: (1)

that test and evaluation were separable functions and could therefore be performed on the same system by different people in independent agencies, and (2) that all data had equal dignity. This marriage meant that anyone with funds, authority, competence, and access to the hardware being developed could collect system data; and anyone else could perform evaluations upon such data. SIDTC spawned the later fragmentation of test and evaluation functions in the Army which was criticized by the General Accounting Office (1984).

The new method consists of effective exploitation of what provisions for test and evaluation do exist in the NDI process. Under the Concept Based Requirements System, there are four prior alternatives which are to be considered before a new materiel acquisition (developmental or NDI) program can be initiated. They are changes to existing tactics, training, doctrine, and organizational structure (AMC/TRADOC, 1987, p. 1.2), all of which are expected to be faster, easier and cheaper than buying a new product. Assuming none of those four will resolve the battlefield deficiency, there are some general steps which must be completed before an NDI acquisition can begin (Figure 1). There are two important things to note. The first is that none of those steps is assigned specifically to an Army test and evaluation agency. The second is that, traditionally, T&E personnel have not participated in "the front end" of the acquisition process in which the system performance requirements were first established.

Step 4 in Figure 1 initially appears to be a T&E function, although the "market investigation" as explained by the AMC/TRADOC Materiel Acquisition Handbook more clearly resembles an intelligence-gathering and analysis function. Several authors (e.g., Johnson, et al., 1988, and Peterson, 1987) have proposed sets of questions the market investigation needs to answer, and

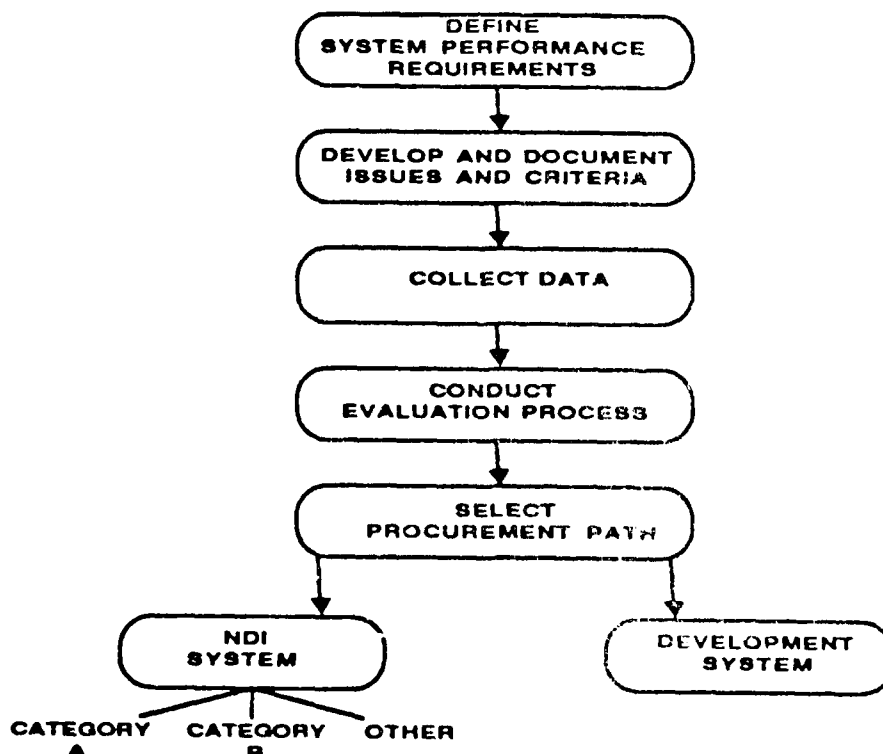


Figure 1. Steps in Determining the Acquisition Approach
(from Johnson, et al., 1988, p. 4-2)

many of them seem to inquire into areas outside the scope of traditional T&E. General Morgan and Dr. Klein suggest (1986, p. 4) that T&E personnel will be invited to participate at appropriate points in the market investigation, and it is here that we believe the bulk of T&E activities in NDI should occur.

If test and evaluation personnel are also permitted to participate in defining the system performance requirements, they can be expected to bring to that process the clarity, precision and objectivity which ought to characterize the conduct of military T&E. The result should be not only an Independent Evaluation Plan which is tied to objective measurement of manned system performance, but performance requirements themselves which are objective expressions of military need and which, therefore, should be intelligible to military arms suppliers who might be able to offer an NDI source.

Assuming that objective requirements for the performance of a manned system were prepared, the next step would be communicating them to industry. Sufficient follow-up questions should also be included so that potential offerors could perform analyses within their companies to determine whether any already-developed products could become legitimate candidates for an NDI acquisition. What the T&E personnel who contribute to that effort have really done is to structure the incoming data from the offerors (the developmental equivalent of having written a Test Design Plan). The remaining activity is to analyze the data provided by offerors to see whether objective system performance criteria appear to have been satisfied.

If NDI is really to operate as speedily as its promoters are predicting (e.g., Buttrey, 1986), there will be little time for offerors to plan, conduct and report results of new tests. Instead, in most cases the Army will have access to some of what data already exist at the offeror's facility. If this amount of information is insufficient to determine whether or not acquisition by NDI is feasible, additional data may be obtained from a subsequent request to the offeror, from an independent laboratory or user, from Army testing, or from early user test and experimentation (see HQDA, 1986).

Application to A Specialty Program

To show how our concept for NDI test and evaluation might work out, we have elected to explain it in the context of what we believe to be the newest of the Army's major programs concerning materiel acquisition: MANPRINT. Although this program has received wide publicity by a number of Army commands and agencies and has been given media attention (including even a lengthy article in the New York Times (Cushman, 1987), we believe that there are currently few good examples of it. MANPRINT is the acronym for "Manpower and Personnel Integration," which seeks to apply to both NDI and developmental programs the latest technology in six "domains" concerning the human component of the system: manpower, personnel, training, human factors, system safety, and health hazards. Here is how we believe our NDI concept of test and evaluation could work in that specialty program:

Clearly, the MANPRINT Joint Working Group (MJWG) (or its designated agent) needs to participate in developing the performance requirements for the manned system (including the three parts required by Army Regulation 602-2: soldier performance standards, training burden, and soldier aptitude). As the O&O Plan nears completion of its first draft, the MJWG selects or adapts (or, less desirably, builds) a simulation model for the system (based upon its class and

mission). This model provides the basis for developing system performance and RAM criteria for each of the functions and subfunctions required to accomplish the mission, and has an architecture which provides for detailing critical operations and maintenance tasks and for accepting actual soldier performance data (in time and accuracy dimensions).

Next, the MJWG begins drafting the target audience description (TAD), which would ideally contain the probable available numbers and types of soldiers, their (ASVAB) aptitude profiles, the probable available training, and the performance of which these soldiers are capable.

Following receipt of specific design information on candidate NDI systems, a version of the simulation model is created for each candidate. Next, the manufacturer of each NDI candidate is queried for time and accuracy data from tests. If none are available, data can be generated from comparability analysis of like tasks on other systems. (This provides, in our view, an incentive for offerors to provide actual performance data on their own systems.)

Exercises of the simulation model with actual or comparability data can be run for each NDI candidate to determine the maximum crew size for operations and maintenance (based on comparisons of the projected number of available soldiers who fit the Target Audience Description with the planned density of the new system in the Army). These runs of the model will also show the likely soldier contribution to manned system effectiveness (as a function of aptitude distributions within the MOSs selected).

Critical tasks which have the greatest impact on manned system effectiveness and those critical tasks which seem to show the greatest difference between NDI candidates can be further analyzed to determine whether hardware, software or training "work-arounds" can be devised, or whether those tasks simply demand high soldier aptitude for correct and timely performance. Such tasks are prime candidates for actual field tests with soldiers meeting the Target Audience Description.

By using simulation and field tests to complement each other, the cost and length of testing can be dramatically reduced; and partial field testing significantly reduces the risks of relying entirely upon the assumptions of simulation. In this example, we have proposed MANPRINT field testing only at those points which prior analysis has shown to be the most crucial to system effectiveness and supportability.

Conclusions

If the goal of placing quality weapons in the hands of our fighting men is to be met, the traditional functions of test and evaluation must still be performed. Although the NDI acquisition concept appears to downplay those functions, we believe there is adequate conceptual leeway for them to be performed--albeit in nontraditional ways. In NDI, much of the government's role shifts from data collection and analysis to verification and corroboration of contractors' and others' data and evaluation of those data. By also assisting in the preparation of precise and objective initial system performance requirements, T&E personnel can expedite the speed and understanding with which all personnel involved in the acquisition perform their various roles.

References

Buttrey, Glen (1986). Acquisition Streamlining in Practice. RD&A Magazine, XXVII, 6, 4-6.

Cushman, John H. Jr. (1987) Making Weapons That Fighting Men Can Use. New York Times, June 21, 1987, Section 3, 1-2.

Headquarters Department of the Army (1986). DACS-ZB MSG, DTG 0103 May 86, subject: Early User Test and Experimentation.

Hightower, Paul E. (1988) NDI and MIL-Spec. Defense Science, 7, 5, 50-51.

Johnson, Kenneth, Riviello, Robert, Rossmeissl, Paul, and Shields, Joyce (1988). Handbook for Nondevelopment Item (NDI) Acquisition. (AMC Pamphlet 602-2). Alexandria, VA: U.S. Army Materiel Command.

Lehnes, Robert R. (1987) NDI: The MSE Acquisition Strategy. RD&A Magazine, 28, 1, 1-5.

Morgan, MG Robert D. and Klein, Ted J. (1986). The Acquisition Method of First Choice. RD&A Magazine 27, 1, 4-6.

Peterson, Gayle D. (1987). NDI at the Belvoir R&D Center. RD&A Magazine, 28, 1, 14-20.

Roche, George (1987). A World Without Heros: The Modern Tragedy. Hillsdale, MI: Hillsdale College Press.

U.S. Army Materiel Command / U.S. Army Training and Doctrine Command, Materiel Acquisition Handbook, AMC/TRADOC Pamphlet 70-2, 26 March 1987.

U.S. General Accounting Office (1984). The Army needs more comprehensive evaluations to make effective use of its weapon system testing. (Report NSIAD 84-40). Washington, DC.

Weinberger, Caspar W. (1987). Quality Leadership and Absolute Integrity. Program Manager, XVI, 5, 2-4.

ERROR IS ALSO A HUMAN FACTOR: TASK DIFFICULTY IN CONTEXT

Carl W. Lickteig
U.S. Army Research Institute
Robert S. Du Bois
Universal Energy Systems, Inc.

This paper reports an interesting, and potentially misleading, instance of differences in performers' task difficulty ratings. In a recent evaluation, operators' experience with an automated system resulted in higher ratings of task difficulty under normal, nonautomated, conditions compared to ratings of control operators who used only a nonautomated system. The effect seems reminiscent of an elder's familiar "When I was a kid ..." accounts of how much more difficult things were in days gone by. We suggest this error is due to the context of automated experience, and may not be uncommon in workload estimation and, particularly, the evaluation of systems designed to reduce operator workload.

A reliance on subjective data, despite the science's benchmark requirement for objective data, is inevitable for the resolution of many Human Factors issues, particularly workload. In a recent examination of the workload concept, Gopher and Donchin (1986) stress its consideration as a hypothetical construct which defies reduction to solely empirical terms and invokes "...processes or entities that are not directly observable." Similarly, Wickens and Yeh (1983) suggest that subjective measures may be more globally sensitive, detecting central processing requirements not tapped by more discrete, empirical measures.

Fortunately, subjective measures of workload, and particularly task difficulty, are not only readily obtainable and intuitively valid, but often strikingly reliable. Gopher and Donchin (1986) review an impressive pattern of subjective workload reliability coefficients, above 0.90, across a wide variety of tasks, input modalities, and response modes. Our intent in this paper is not to undermine the validity or utility of subjective measures, but to alert the reader to a potentially important contextual factor in their obtainment.

Method

As part of a recent simulation-based evaluation of the usefulness of automated navigational systems for small-unit armor missions, sixty tank commanders rated the difficulty of performing nine basic navigational tasks. Initially, tank crew and platoon participants were randomly assigned to one of the following tank simulator, experimental conditions: (a) a control condition with only conventional paper map, protractor, and grease pencil navigational aids; (b) a grid condition with an automated navigational system that provided a steer-to display for the

driver, and for the commander, a grid map display of the battlefield with a vehicle icon continuously updated as the vehicle moved; and, (c) a digital condition identical to the grid condition, except for the addition of terrain features such as roads, trees, and rivers onto the commander's map display.

Following an initial day of classroom and simulator training, participants spent two days performing a series of simulation-based road marches and platoon combat missions. After finishing these exercises, tank commanders completed several subjective measures including the task difficulty ratings which are the focus of this paper. Participants' estimates of task difficulty were included as auxiliary measures of the navigational system's potential for reducing the difficulty of performing land navigation tasks. For a more complete discussion of the methodology and findings of this evaluation the reader is referred to Du Bois and Smith (in press).

All tank commanders rated the difficulty of nine basic land navigation tasks on a 7-point rating scale ranging from extremely easy (1) to extremely difficult (7). Each commander completed two different task difficulty ratings for each task: (a) the difficulty of the task as performed in the experimental condition assigned--control, grid, or digital--in an M1 tank simulator, and (b) the difficulty of performing the same task in an actual M1 tank not equipped with an automated navigational system. For estimating task difficulty in an actual M1 tank operating in the field, participants were instructed to base their ratings on their prior experience with tanks, an average of 6.5 years. The order of the tasks, question blocks (simulator first and tank second), and response options were fixed.

Results

The general findings from the overall evaluation are summarized before presenting the data on task difficulty ratings. To assess the adequacy of the randomized assignment procedures for the among subjects design, data were collected on eight different soldier measures including Armor experience, aptitude scores, and a land navigation skills test. Group equivalence was obtained on all measures. Overall, crews and platoons equipped with either the digital or grid automated navigational system outperformed the control group on 32 of 36 dependent measures including: number of missions and fragmentary orders successfully completed, time required, distance traveled, fuel expended, and speed and accuracy of reporting both own tank and enemy target locations.

Participants' estimates of task difficulty for their simulation-based performance parallel the above findings on objective performance. Commanders operating with either of the automated navigational systems rated task performance on eight of the nine land navigation tasks significantly easier than commanders in the control group. Their mean difficulty rating for these tasks was 2.25 compared to the control group's mean rating of 4.18,

$t(57) = 6.57, p < .001$. The only simulation-based task not rated as significantly easier, maintaining platoon formations, is discussed in a later section.

For tank-based difficulty ratings, however, commanders who had used the automated navigational systems rated seven of the nine tasks as significantly more difficult to perform in an actual M1 tank than participants from the control group (see Table 1 and Figure 1). Consistent with this pattern of differences, within group comparisons indicated that commanders in the automated groups rated simulation-based task performance as significantly easier than tank performance, and the control group rated tank performance easier than simulation-based performance. In summary, equivalent and experienced tank commanders provided significantly nonequivalent estimates on the difficulty of performing land navigation tasks in an actual tank in the normal operational setting, after performing the tasks under either automated or nonautomated conditions in a simulated setting.

Table 1

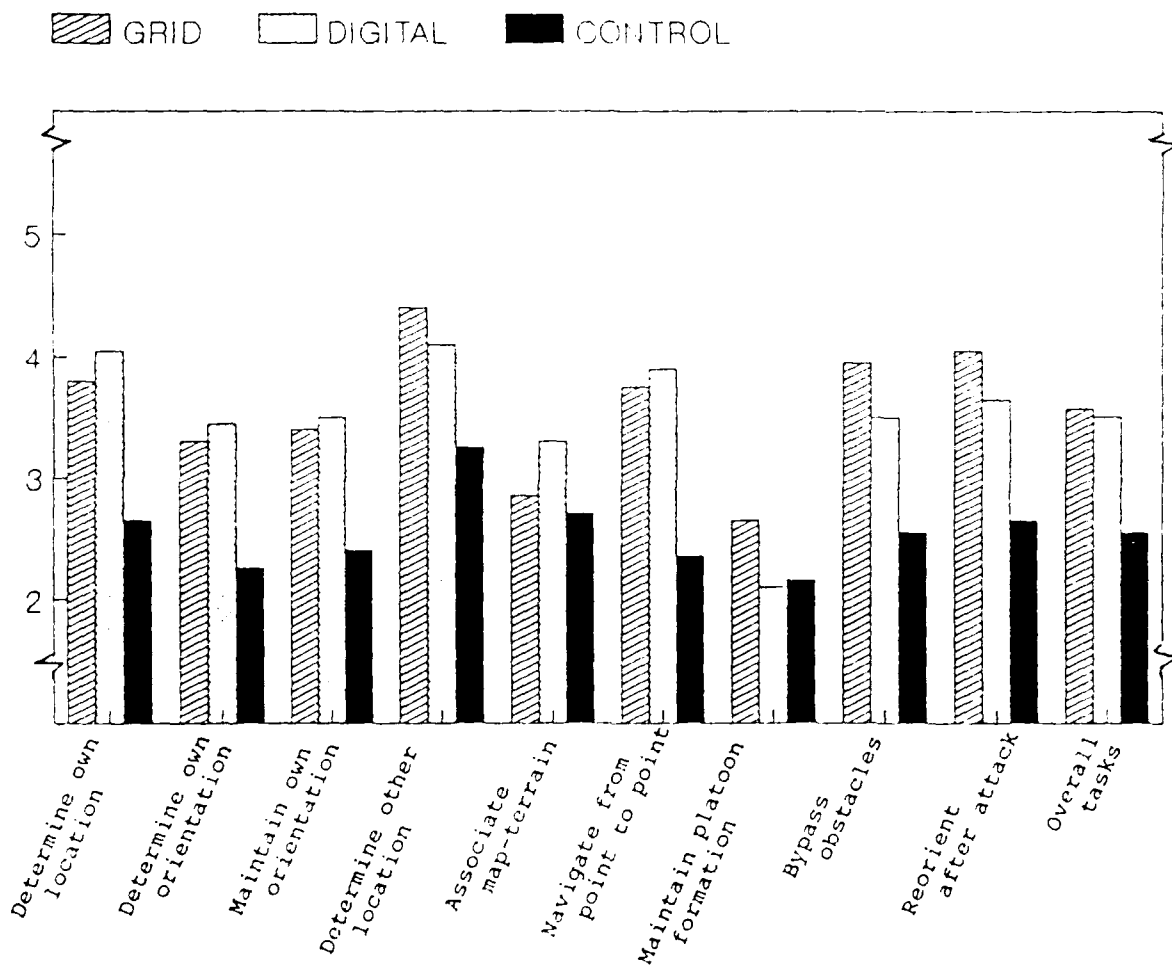
Task Difficulty Data for Tank-Based Performance:
Mean, Standard Deviation (SD), Planned Comparison
of Grid and Digital Conditions versus Control

Task	Mean (SD) By Condition			Planned Comparison	
	Grid	Digital	Control	I-Value	p
1. Determine own location	3.80 (1.06)	4.05 (1.54)	2.65 (1.18)	-3.651	.001
2. Determine own orientation	3.30 (1.34)	3.45 (1.73)	2.25 (1.16)	-2.868	.006
3. Maintain own orientation	3.40 (1.50)	3.50 (1.64)	2.40 (0.99)	-2.728	.008
4. Determine other location	4.40 (1.50)	4.10 (1.65)	3.25 (1.33)	-2.433	.018
5. Associate map-terrain	2.85 (1.27)	3.30 (1.63)	2.70 (1.34)	-0.964	.339
6. Navigate from point to point	3.75 (1.77)	3.90 (1.52)	2.35 (1.14)	-3.593	.001
7. Maintain platoon formation	2.65 (1.76)	2.10 (1.59)	2.15 (1.42)	-0.515	.608
8. Bypass obstacles	3.95 (1.70)	3.50 (1.54)	2.55 (1.19)	-2.876	.006
9. Reorient after attack	4.05 (1.64)	3.65 (2.13)	2.65 (0.99)	-2.648	.010
Overall tasks	3.57 (0.90)	3.51 (1.22)	2.55 (0.71)	-3.730	.000

Note: Task difficulty was rated on a seven-point scale (i.e., 1 = extremely easy, 2 = quite easy, 3 = slightly easy, 4 = neither easy nor difficult, 5 = slightly difficult, 6 = quite difficult, 7 = extremely difficult).

Figure 1

Task Difficulty Ratings by Test Condition



Note: Task difficulty was rated on a seven-point scale (i.e., 1 = extremely easy, 2 = quite easy, 3 = slightly easy, 4 = neither easy nor difficult, 5 = slightly difficult, 6 = quite difficult, 7 = extremely difficult).

Discussion

A potential explanation for these differences in task difficulty ratings for actual tank performance might include response biases caused by questionnaire artifacts, such as the fixed order of tasks, question blocks, and response options. A perusal of the findings argues against these interpretations. First, the pattern of differences is not consistent across all task ratings. Two of the nine items, the fifth and seventh, show no significant differences. The location of these items, in the middle of the series, suggests that the pattern of differences was not the result of questionnaire induced response sets.

Second, the content of these two items indicates that participants were responding thoughtfully, in the context of the actual simulation capabilities recently experienced. As with any low cost simulation, the Simulation Networking (SIMNET) test bed used in this evaluation, entails certain fidelity limitations (DuBois & Smith, in press). The SIMNET computer generated imagery, for example, provides a less differentiated set of terrain features than most natural terrain settings. This degraded detail makes performance of map-terrain association, the fifth item, more difficult in the simulator than in an actual tank. Similarly, the commander's rotatable cupola in the simulator affords a limited, horizontal field-of-view compared to the widely dispersed vision blocks of the M1 tank. Maintaining platoon formations, the seventh item, requires visually monitoring the relative location of the platoon's four tanks.

The effect of differential simulation contexts, non-automated versus automated task performance, is reflected in commanders' difficulty ratings for actual tank performance. The experience of performing tasks made more difficult by low-fidelity simulation and unaided by the automated system, resulted in relatively easy ratings of task performance in an actual tank. In contrast, the experience of executing tasks with the aid of a system that directly supported task performance, resulted in more difficult ratings of nonautomated task performance in an actual M1 tank.

Various comparative judgement theories might account for this effect. Helson's (1964) Adaptation Level Theory, for example, suggests the shift in ratings may be due to the effects of recent and past experience, background and residual stimuli. To counter response bias explanations of the effect, a current research effort is counterbalancing the order of question blocks to control for background stimulus effects, and obtaining pretest measures of tank-based task difficulty, residual stimuli, prior to simulation-based performance. This work will attempt to replicate the reported effect on navigational tasks, and assess generalization of the effect to a new set of tasks--command, control and communication.

Given the well-founded reliance of workload assessment on subjective measures, the present finding suggests that researchers involved in automated system design and workload reduction efforts should attend more carefully to the effect that operator exposure to such systems might have on estimates of task difficulty and related measures of perceived workload. Experience with automated systems may, in fact, generate more valid workload estimates by broadening the rater's comparative base and scale range. This may prove particularly important in settings such as the military where social mores emphasizing competition and competence tend to restrict estimation of one's job requirements as "too difficult."

References

- Derrick, W. L. (1988). Dimensions of operator workload. Human Factors, 30 (1), 95-110.
- Du Bois, R. S. & Smith, P. G. (in press). The effect of position navigation (POSNAV) information displays on the performance of armor crews and platoons. Army Research Institute, Technical Report.
- Ericcson and Simon (1980). Verbal reports as data. Psychological Review, 87, 215-251.
- Ericcson & Simon (1984). Protocol analysis: Verbal reports as data. Cambridge, Mass: Massachusetts Institute of Technology Press.
- Gopher, D. & Donchin, M. (1986). In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), Handbook of Perception and Human Performance: Vol 2. Cognitive Processes and Performance. (pp. 41-1 to 41-49). New York, NY: John Wiley and Sons.
- Helson, H. (1964). Adaptation Level Theory. New York: Harper & Row.
- Nisbett & Wilson (1979). Telling More than we can know: Valid reports on mental processes. Psychological Review, 84, 231-259.
- Wickens, C. D., & Yeh, Y.Y. (1983). The dissociation of subjective ratings and performance: A multiple resources approach. Proceedings of The Human Factors Society Twenty-Seventh Annual Meeting, 244-248.

RESULTS OF A SEARCH FOR COGNITIVE SKILLS

John A. Modrick

System and Research Center, Honeywell, Inc.

The form of the ideas presented can be described generously as inchoate and embryonic. Motivated by a reaction of alarm and outrage over the state of the art, we have been working to structure the problem of defining cognitive skills while developing a working approach in parallel. It is difficult in this paper not to encounter the risk of offending persons whose work is critized often out of context and against criteria or purposes other than those of the originators. Our intent is to evaluate the state of the art rather than personal attack and we hope that our apologies will not be needed. This work has been supported by a combination of internal funding and personal time.

THE PROBLEM: TAXONOMY OF COGNITIVE SKILLS

The original objective of this project begun in 1981 was to compile a listing or taxonomy of cognitive skills: a set of terms for describing or classifying the activities of people in situations where information processing rather than psychomotor control is the primary operation. The taxonomy was to be used subsequently in a method for the decomposition of a protocol of operational activities into the component skills and abilities which made up the operational protocol. These protocols would consist of the multi-task, time sharing activities of a tank commander or crew, pilot of a helicopter gunship or brigade operations officer, for example. Front end analysis of missions for defining system concepts and designs are a long range objective. The original approach to compiling the taxonomy was collation of categories of skills, techniques and principles from the technical literature in psychology, human factors, cognitive science, management science and artificial intelligence.

There have been many classifications of behaviors, skills and abilities produced for various purposes in personnel, psychometrics and human factors. Among the more familiar as those used for task analysis and the outcomes of factor analytic studies such as Guilford's structure of intellect. Many have been useful in application to practical applied problems. The book "Taxonomies of Human Performance" (Fleishman and Quaintance, 1984) provides a comprehensive survey both approaches and schemes for classification of skills and abilities.

Human activities during the period when these classifications were developed were more heavily loaded with psychomotor, sensory and perceptual factors and the separation between human and machine functions was sharper. However, there have been increases in recent years in the amount of automation in manned systems and the human-system relationships. The role of the human had become more cognitive than sensory or manual control; it was becoming a role of supervisory control of automated, intelligent systems requiring more knowledge rather than psychomotor skill. The human operator in the future will perform activities that are less well structured and more complex under conditions of partial knowledge and uncertainty such as planning at tactical and strategic levels, monitoring the execution of planned actions for deviation from expected course of events, goal setting, recognition of patterns of events, interpretation of tactical information and

analyses, detection of infrequently occurring events and redirection of a plan in response to changed operational conditions and emergencies. These activities have always been the responsibility of the human. However, little effort has been given to support them and we have relied on someone who has a "knack" for them tempered and matured by experience. These activities are becoming commonplace for the typical individual and too critical to leave to happenstance. In the past the user or operator of a system was very busy directly controlling actions of platforms. Now the platforms are capable of operating autonomously once they have been committed. The airplane can fly itself and weapons can acquire targets and fire. The user's problem is to use these systems to fight. Commanders at all levels are no longer needed to control engagements but to plan when, where and how he wants to engage.

DISAPPOINTING RESULTS

The results were very disappointing. There is a complete lack of terminology, classifications or taxonomies of cognitive skills suitable for human factors research and engineering that be can considered well-defined, systematic, validated and comprehensive. There is a plethora of "ad hoc" lists and classifications designed for the purposes of a specific study. Most discussion of cognitive activities and skills are at a very high level in terms such as planning, problem solving, decision making, etc. with little specificity or even definition of terminology.

There is a big jump to further levels of resolution going first to conventional task analysis and then to what I call psychometric abilities. Review and analysis of task analytic methods (Bennett, 1971; Christman, 1977; and Companion and Corso, 1983) reveals several deficiencies and little that is adequate. The terminology of task analysis seems to fit best in the perceptual and motor areas. The classification by Berliner, Angell and Shearer (1964) is representative. Mediational processes is the most cognitive category. It is one of four processes; the others are perceptual, communication and motor. Mediational is broken down into two activities with the following specific behaviors:

- Information Processing: categorize, calculate, code, compute, interpolate, itemize, tabulate and translate.
- Problem Solving and Decision Making: analyze, calculate, choose, compare, compute, estimate and plan.

The first reaction to these lists of behaviors is that today they are executed in part or whole by computers supporting the human. This level is also far too specific. We need intermediate categories within the activities of information processing, problem solving and decision making which deal with what, how and why something is being analyzed, computed, etc. Similarly, the processes should be further differentiated into application domains and operational goals. The specific behavioral skills are fundamental building block which are put together in various combinations to accomplish higher level goals.

As these changes in the human role, human-system transactions and human-system interface have

been occurring, there has been a need for changes in task analytic methods. The need has not yet been met. Card et al. (1983) have developed the model called Goals, Operators, Methods and Selection (GOMS) for description and analysis of activities in manuscript editing and human-computer interactions. Bond et al. (1983) developed a method for cognitive task analysis based on the GOMS approach.

Identification and analysis of psychometric abilities is an active area of research but it is even more molecular and therefore harder to relate to operational activities. This work can be illustrated by Sternberg (1977) on solving analogies, Carroll (1981) on cognitive abilities, Feuerstein et al. (1979) on perceptual abilities and Hunt (1983) on the nature of intelligence. Hogan et al. (1987) have identified a battery of twenty-four test instruments for use in classifying training tasks.

Experimental psychologist and cognitive scientists do no better; the necessity of using naive college sophomores as subjects and the drive to simplify, standardize and control lead to tasks such as solving geometry problems (Greeno), planning household errands (Hayes-Roth, 1979), choosing between wagers, physics problems and the ever present water bottle problems. Fischhoff and Beyth-Marom (1983) in a study of bias in evaluation of Bayesian hypothesis use the "tasks" of hypothesis formation, assessing component probabilities, assessing prior odds, assessing likelihood ratio, aggregation, information search and action. The absence of process models for operational activities, the tasks in experiments and the cognitive abilities make generalization among them impossible.

I have collected classifications and articles on specific cognitive activities and I now have a large collection. Unfortunately, they seem to be chaotic rather than orderly; they are also heterogeneous and contradictory.

A CONSTRUCTIVE RECOMMENDATION

When faced with chaos or intractable complexity, the creative scientist simplifies and imposes structure. We will outline an approach and conceptual framework to provide a proper taxonomic system. A diagram of the approach is presented in Figure 1.

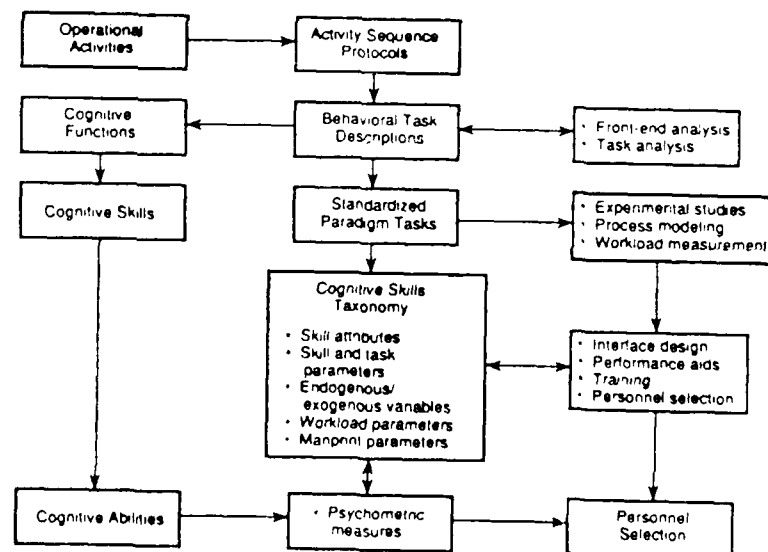


Figure 1 Approach to Developing a Cognitive Skills Taxonomy

Compile an Inventory of Operational Activities of People in Some Specific Domain

We must begin by collecting Activity Sequence Protocols of the operational activities of people and analyzing the activities into the component cognitive functions, processes, tasks, skill/knowledge and abilities. It is not feasible or useful to try to compile a list that encompasses the entire spectrum of human activity. Instead we must limit the inventory to a tractable domain, perhaps the mission of a howitzer or attack helicopter or even a segment of the mission and specific operational objectives. The activity sequence will be broken down into component tasks and functions. Behavioral task descriptions will be prepared and cognitive functions will be abstracted from them. Cognitive skills will be derived from the functions.

Prepare Behavioral Task Descriptions

The purpose of Behavioral Task Descriptions is to capture the operational context for the activities. The behavioral task descriptions will consist of the sequence of actions in exercising the skill in a given operational context including the stimuli, cues or information responded to; responses made and the control activations in making the response; and the cognitive operations performed in mediating between cuing conditions and output response. It is important to identify the non-observable "in the head" responses characteristic of many cognitive activities.

Hierarchical Structure That Spans the Range of Real World Operational Activities of People

We must impose a vertical echelon structure that spans the echelons of planning, command and control. There are two approaches on which we can build to address the structure of vertical echelons. Anthony (1981) differentiates among three levels of administration:

- Strategic Planning: Establish long term objectives and use of resources;
- Management Control: Formulate, implement and monitor programs to achieve strategic goals;
- Operational Control: Day-to-day operations management of production or service.

Wohl (1981) offers a related paradigm for AF tactical C2 in terms of pre- and inflight levels which deal with planning and commitment vs control and coordination, respectively. Time, amount of information and aggregation of information decrease from pre- to inflight.

Our assumption at this time is that the cognitive skills are the same at each hierarchical level. Some of the tasks may be different but the principal differences will be in the time available to act and the scope and aggregation of the data.

Select or Construct a Standardized Task For Each Skill Category

At this point we would select which will serve as the paradigm for that skill. We will be able to use these tasks for off-line, laboratory research and analysis to develop task and process models, make workload estimates under specified conditions and develop the kinds of information on human

resource demands needed by MANPRINT types of programs and system engineering. Again, existing batteries for workload measurement and used in studies by Hart, Wierwille, Wickens and Donchin can be used provisionally to get started. They can be standardized laboratory paradigms or standardized job tests.

Taxonomy of Cognitive Skills

The cognitive skill will then be organized into a taxonomy. The initial cognitive terminology would undoubtedly be commonly used terms and concepts. There are several sources of function/task data bases that would provide a provisional classification scheme, such as Thompson et al. analysis of intelligence activities, the GOMS model (Card et al., 1983), DARPA/NOSC battle management and the US Army work on brigade C3 at Fort Leavenworth. The frequency data from decomposition of the operational activities would then be subjected to some kind of **cluster analysis** to prune and dimensionalize the skills inventory. Bennett, for example, did a factor analysis on twenty-five verbs of the type used for task analysis; the verbs were used to make judgement about ten tasks. The factor analysis yielded four factors on which the verb clustered: cognitive, social, procedural and physical.

Each cell would consist of an independent skill unit, a definition of the cognitive operations, associated data of variables that affect it and all the good human factors information on requirements for interface design, support and training.

Cognitive Ability Level

The final step is to relate the cognitive skills and paradigmatic tasks to abilities, aptitudes and their instruments.

CONCLUSION

The approach proposed is a big undertaking but it is necessary. Better utilization of human abilities is critical to realizing the potential of automated, intelligent systems. There are no shortcuts: technology is not going to actualize itself. Our technology for effective design must be based on a sound psychometric, experimental and theoretical foundation. It is a time-consuming, nasty job but someone has to do it.

REFERENCES

- Anthony, R. N. Planning and Control Systems: A Framework for Analysis. Graduate School of Business Administration, Harvard University, Boston, MA, 1965.
- Berliner, C. Angell, D. and Shearer, J. W. Behaviors, Measures and Instruments for Performance Evaluation in Simulated Environments. Presented at Symposium on Quantification of Human Performance, 17-19 August 1964, Albuquerque, NM, 1964.
- Bennett, C. A. Toward Empirical, Practicable, Comprehensive Task Taxonomy. Human Factors, 1971, 13, 229-235.
- Bond, L. Eastman, R. Gitomer, D. Glaser, R. and Lesgold, A. Cognitive Task Analysis of Tech-

nical Skills: Rationale and Approach. Learning Research and Development Center, University of Pittsburgh, Pittsburgh, PA, 1983.

Card, S. K. Moran, T. P. and Newell, A. The Psychology of Human-Computer Interaction. Lawrence Erlbaum Associates, Hillsdale, NJ, 1983.

Carroll, J. B. Ability and Task Difficulty in Cognitive Psychology. Educational Researcher, 1981, 11-21.

Christman, N. J. A Human Factors Taxonomy. Thesis for the Degree of Master of Military Art and Science, U. S. Army Command and General Staff College, Fort Leavenworth, KS, 1977.

Companion, M. A. and Corso, G. M. Task Taxonomies: A Critical Review and Evaluation. International Journal of Man Machine Sciences, 1983.

Feuerstein, R. Rand, Y. Miller, R. and Hoffman, N. Instrumental Enrichment. University Park Press, Baltimore, MD, 1979.

Fischhoff, B. and Beyth-Marom, R. Hypothesis Evaluation From a Bayesian Perspective. Psychological Review, 1983, 90, 239-260.

Hayes-Roth, B. and Hayes-Roth, F. A. A Cognitive Model of Planning. Cognitive Science, 1979, 3, 275-310.

Hogan, J. Broach, D. and Salas, E. A Task Information Taxonomy for Instructional System Design. Naval Training Systems Center, Orlando, FL 32813-7100, 1987.

Sternberg, R. Intelligence, Information Processing and Analogical Reasoning. Erlbaum, Hillsdale, NJ, 1977.

Thompson, J. H. Hopf-Weichel, R. and , ARI TEchnical Report.

Wohl, J.D. Force Management Decision Making for Air Force Tactical Command and Control. IEEE Transactions on Systems, Man, and Cybernetics, SMC-11, 1981, 618- .

November 28, 1988

Applications of the Cognitive Requirements Model

Paul G. Rossmeissl
Jan Charles

Hay Systems, Inc
2000 M Street NW
Washington DC. 20036

An accurate measurement of the mental difficulty of a job or task is a key element in the successful planning and management of human resources. HAY Systems, Inc. has developed a procedure, called the Cognitive Requirements Model or CRM, to estimate the cognitive demands of human task performance in a manner that permits provides useful information to human resource managers.

The cognitive requirements model was developed through a careful review of the literature on human information processing theory and cognitive models (eg., Anderson, 1981, 1983; Sternberg, 1977; Snow, Fedrico & Montague, 1980). This research led to the identification of twelve factors which are of particular importance in determining job or task difficulty. These factors became the model's evaluation criteria and are listed in Table 1.

Table 1
Cognitive Requirements Model
Evaluation Criteria

<u>Cognitive Factors</u>	<u>Moderator Factors</u>
Working Memory (WKM)	Psychomotor (PSY)
Quantity of Data (QNT)	Long Term Memory (LTM)
Multiple Processing (MLT)	Task Repetition (REP)
Data Interpretation (INT)	Perceived Risk (PRK)
Problem Solving (PBL)	Time (TME)
Decision Mapping (DMP)	Environmental Factors (ENV)

The first six factors are considered to be of direct significance to the cognitive difficulty of a task. The remaining factors are moderator variables that affect cognitive task requirements in an indirect fashion. For example, two tasks may have the same scores on the cognitive factors, but one must be accomplished in half the time as the other. The task that must be accomplished quicker will likely be more difficult for humans to perform.

Each of the factors within the model is evaluated on a five step rating scale, similar to those used by Schufeldt and Layton (1972) in evaluating the mental workload of aircraft pilot tasks. Each rating scale is anchored on observable and measurable performance requirements or established hierarchical structures of mental complexity.

The data from the evaluation the rating scales are entered into the model's evaluation algorithm that is highly interactive in form. This form was required to capture the interrelationships among the cognitive and moderator variables that have been documented in the literature (eg. Anderson, 1983; Snow, Fedrico & Montague, 1980). The output of the evaluation process is a single number (the cognitive loading score) which indexes the overall cognitive complexity of the job or task.

This paper reports on the initial study of the utility of the cognitive requirements as a human resource tool. Of particular concern was the investigation of the model's psychometric properties and the validity of its measurements.

CRM PSYCHOMETRIC PROPERTIES

An investigation into the psychometric properties of the CRM was undertaken to accomplish two major objectives. The first of these objectives entailed an analysis of the structure and distributional properties of the model. The second major objective in evaluating the psychometric properties of the CRM was the determination of the model's inter-rater reliability.

Method

Tasks. In order to investigate the psychometric characteristics of the CRM, a sample of 131 tasks were evaluated with the model. All of the tasks entailed either U.S. Army engine, aviation, or vehicle maintenance procedures that were likely to be performed by a skill level one (entry level) soldier. In order to estimate the inter-rater reliability of the model, a subset of thirteen helicopter engine maintenance tasks were selected from the total set of 131 tasks. These tasks were to be rated by more than one analyst.

Procedure. Each task was evaluated by analyst highly trained in the use of the CRM. Three different analysts were used to evaluate the complete sample 131 tasks. These tasks were divided equally among the three analysts. The thirteen tasks that were to serve as the base for estimating the inter-rater reliability or agreement of the CRM procedure were evaluated by all three analysts.

Results

Analysis of evaluation scale values. Summary statistics for the twelve CRM evaluation scales are presented in Table 2. With the exception of one dimension (Environmental Conditions) each scale showed sizable variation across the tasks under evaluation.

Table 2
Summary Statistics for CRM Scale Values

Scale	Mean	S.D.	Range
Working Memory	2.0	1.0	4.0
Quantity of Data	3.0	1.2	4.0
Multiple Processing	1.1	.2	1.0
Data Interpretation	1.8	.7	2.0
Problem Solving	1.5	.8	3.0
Decision Mapping	2.1	.4	3.0
Psychomotor	2.3	.4	2.0
Long Term Memory	2.8	.6	1.5
Task Repetition	4.3	1.2	4.0
Perceived Risk	2.1	1.1	3.0
Time Constraints	1.2	.7	3.0
Environmental Cond.	3.0	0.0	0.0

Tables 3 and 4 show the inter-correlations among the scale values for the CRM's cognitive and moderator factors. Using the r to z transformation (Pearson & Hartley, 1966), the critical value for statistical significance ($p = .05$) for these data is .17.

Table 3
Inter-correlations among CRM Scale Values
Cognitive Factors

	QNT	MLT	INT	PBL	DMP
WKM	.44	.53	-.04	.56	.15
QNT		.23	.35	.41	.38
MLT			-.07	.47	.24
INT				.05	.37
PBL					.19

Analysis of inter-rater reliability . To assess the inter-rater reliability of the model's output the intra-class correlation coefficient (ICC) (Nunnally, 1978) was calculated. The ICC for the three analysts who all rated the subset of thirteen engine maintenance tasks was .81 indicating strong inter-rater agreement.

Table 4
Inter-correlations among CRM Scale Values
Moderator Factors

	LTM	REP	PRK	TME	ENV
PSY	-.11	-.15	.16	-.04	K
LTM		-.15	-.07	.13	K
REP			.12	-.13	K
PRK				.08	K
TME					K

Discussion

Overall, the CRM was found to exhibit favorable psychometric properties. The task difficulty indices produced by the model were reliable across users and showed sufficient variance that floor or ceiling effects when applying the model should be minimal.

The examination of the inter-correlations of the evaluation scales indicated that the use of twelve scales seems to be appropriate. None of the inter-correlations among the moderator variables was found to be statistically significant, suggesting that each makes a unique contribution to the model. Scale correlations were higher among the cognitive factors, but this finding is not surprising in light of the strong correlations typically found among these factors in cognitive tests (e.g. Hunt, Frost, & Lunneborg, 1973; Pellegrino & Glaser, 1979).

VALIDITY TEST

The ultimate value or utility of any psychometric instrument or procedure is a function of its validity. One source of true task difficulty data can be derived from the judgments of subject matter experts (SMEs). If SMEs possess extensive familiarity with the tasks and with the historical difficulty inherent in their learning or performance, SME judgments would constitute an acceptable basis against which to measure the validity of the task difficulty estimates

Method

Subject Matter Experts. Thirteen instructors at the U. S. Army Transportation Center and School (Aviation Logistics School) served as subject matter experts or SMEs for evaluation of task learning difficulty. Seven senior sergeants at the U. S. Army base at Ft. Hood, TX. were used as SMEs for the ratings of task performance difficulty. All of the SMEs had extensive experience with engine maintenance tasks and knowledge of the capabilities of the soldiers who are asked to learn and perform the tasks.

Materials. The same thirteen turbine engine maintenance tasks that served as the data source for the assessment of inter-rater reliability were used as the basis for the validity data collection. Special data collection forms were constructed to allow the SMEs to evaluate how difficult it would be for personnel within the aviation maintenance specialty to learn and perform the thirteen tasks. Each form contained a seven point difficulty rating scale (seven being the most difficult) and instructions on how to make the ratings. Preceding each task rating scale was a description of the task. The task descriptions were taken from the official technical manual describing the proper procedures for task performance. In addition to verbally describing the activities entailed in performing the task, each description included drawings showing the locations of the relevant engine components.

Procedure. All of the SMEs evaluations were conducted in two group sessions (one session for evaluations of learning difficulty, one session for evaluations of performance difficulty). Each SME was first given a booklet containing the task descriptions and rating scales. They were then asked to read the descriptions and based upon their experience with the tasks rate along the seven point scale the learning or performance difficulty of the described task.

After the SMEs had completed the task ratings, they were asked to rank order the thirteen tasks based upon their difficulty. The easiest task was to receive a rank of one and the most difficult a rank of thirteen. The SMEs were further instructed that two different tasks could not receive the same rank (no ties were allowed).

Results

The results of the correlational analysis of CRM predictions of task difficulty and the SME judgments of learning and performance difficulty for the same tasks are presented in Table 6. The data shown in Table 5 indicate that the CRM is a very good

Table 5
CRM Validity Estimates

Predicted Criterion Measure	r	r ²
Learning Difficulty Ratings	.77	.59
Learning Difficulty Rankings	.79	.62
Performance Difficulty Ratings	.87	.76
Performance Difficulty Rankings	.71	.50

predictor of both learning and performance difficulty. All of the observed validity indices were quite large and statistically significant ($p < .001$).

Discussion

The results in Tables 5 are convincing evidence that the CRM is a valid measure of the difficulty individuals may experience in learning and performing human/machine tasks. The SMEs who provided the criterion data have had sufficient experience with human capabilities and the sampled tasks to permit accurate evaluations of task difficulty. The cognitive loading scores output from the CRM were able to predict the task difficulty criteria with considerable accuracy. Since the CRM does not contain special assumptions for maintenance tasks, the literature on meta-analysis (e.g. Hunter, Schmidt, & Jackson, 1982) indicates that it is reasonable to accept that CRM outputs would be equally predictive of difficulty measures for most human tasks.

Given the strong statistical validity of the CRM as a predictor of task difficulty, it can offer an important contribution to the determination of human task requirements in situations where subject matter expertise with respect to human performance is not available. This capability can be particularly useful for the cognitive evaluation of tasks that are part of a new system design. Since the CRM relies only upon task descriptions for exogenously supplied inputs, it is capable of evaluating tasks based upon information available relatively early in the design process, such as design drawings, human factors task analyses, or prototypes.

In addition to the ability to predict task difficulty for systems that are conceptual or developmental in nature, a CRM analysis can provide valuable information regarding the probable cause (or causes) of that difficulty through evaluation of the exogenous factors that led to the scale values from which it derives its prediction. Isolation of these elements of a task that contribute to its difficulty permits their address through special training, job aids, or re-design of the appropriate hardware or software, hence contributing to enhanced total system performance.

General Discussion

This initial investigation of the cognitive requirements model or CRM has shown it to be a promising tool for addressing the problems of task complexity and human performance. The analyses of variances and correlations of the model's evaluation scale values and the distribution final cognitive loading scores all indicated that the model is psychometrically

sound and consistent. When users of the model are well trained in its procedures, there is substantial agreement among the scores that they produce. Finally, the model's indices of task difficulty were found to be valid when compared to expert decisions on task performance and learning difficulty.

The ultimate test of the CRM, like any other tool or procedure, will rest upon its applications. The model must have utility for applied situations and problems if it is to be of value. In this regard, we believe that there are a number of areas where the model could prove to be useful. The key feature of the CRM in all of these applications rests in its capability to estimate the difficulty of a system's human tasks before the hardware and software of the technology is actually implemented.

One possible application is the use of the CRM to determine the likelihood that a new system can be successfully operated or maintained by the same personnel who are working on the current or predecessor system. The CRM could be used to perform a baseline analysis of the difficulty of the human tasks that are part of the current system. A subsequent CRM analysis, based upon preliminary design information, of the new system's tasks would reveal their predicted difficulty. The human task difficulty of the old and new systems could then be compared on a common metric. If the new system cognitive loading scores are no greater than those of the predecessor system, the same personnel could be expected to operate or maintain the new system successfully (given the reasonable assumption that these personnel are performing satisfactorily on the predecessor system). If the new system's tasks are predicted to be more difficult than desired, the CRM analysis can be used to pinpoint the design factors that led to this difficulty. These factors could then be altered or provided with special job or training aids in order to bring the task difficulty within the targeted range. All of the adjustments to the new system that are based upon CRM analyses could be performed before the system becomes operational when such changes would be very costly.

Another application of the CRM could be to assist in the allocation of functions or tasks between men and machines. Currently, the decision of which aspects of a system to automate is often based on which functions are the easiest to design or engineer the automation. A CRM analysis could determine which functions within a system would likely be most difficult for humans to perform. These functions would then be strong candidates for automation. On the other hand, if a system function would be easy for a human to perform, expensive automation may not be cost effective.

Applications of the CRM of this sort would be particularly useful for organizations with constrained work-forces or training resources. Such organizations would include in particular the military and companies with strong labor union commitments. We are currently using the model in a number of these situations and expect success.

SOLDIER FEEDBACK FOR BETTER PRODUCTS

BARBARA A. JEZIOR
LAWRENCE E. SYMINGTON
CHARLES A. GREENE
U.S. ARMY NATICK RESEARCH, DEVELOPMENT AND ENGINEERING CENTER
AND
ANNETTE M. SALVATO
GEO-CENTERS, INC.

INTRODUCTION

Although the military has been much more responsive than the civilian world to the need for user testing and the application of other Human Factors precepts, its efforts until recently have been primarily centered in a few major systems and product areas. This situation is changing. Programs such as Manpower and Personnel Integration (MANPRINT) now mandate the integration of the soldier-user in the design loop for all products in all phases of the development process.

Although weapons and aircraft have been the attention getters in design issues, Natick's products -- rations, environmental protection items, clothing, shelters, and airdrop systems -- also demand an uncompromising application of human factors principles. One could argue that since Natick's products center on individual soldier life support it is even more critical that we target the soldier's physiological and psychological parameters in order to assure soldier use and acceptability of our products.

Acceptability is one of the critical human factors issues for Natick items, as a soldier simply will not use an item he or she finds unacceptable. An unused item equates to the soldier lacking a battlefield advantage and wasted money. The costs of Natick's items startle those who have only heard the litany of weapons dollars. For instance, the 1987 procurements for just the Meal, Ready-to-Eat, Battle Dress Uniform, Combat Boot, Jungle Boot, and the General Purpose Medium Tent amounted to \$395 million -- and those items are usually procured annually.

Over four years ago Natick responded to the prevailing human factors climate that pressed for broad user (soldier) interface in product assessment by launching OFIG (Operational Forces Interface Group).

Although civilian and military product developers are aware they must user-test, there are few available models. While much literature exists on research design, there is little that offers a "cook book" schema for implementing a structured program in an organizational setting.

OFIG has evolved a strategy to obtain structured user feedback. We would like to describe the OFIG system and discuss our solutions to some of the compromises inherent in field testing.

OPERATIONAL FORCES INTERFACE GROUP - A CASE STUDY

Organization and Personnel. Natick's structure and mission allowed for an easy birth of a user product assessment program. Initially, Natick assigned an equipment specialist and an Infantry officer from its operations research staff and a human factors psychologist and two technicians from the behavioral sciences staff. This provided a blend of research and product expertise, plus a military member to help access the military user. Program personnel also had ready access to product project officers and a staff statistician. The number of personnel has grown to seven staffers as the program expanded. Organizationally, the OFIG members collaborate closely, but are responsible to their respective management chains.

User Surveys. OFIG surveys soldiers in the United States and overseas on Natick's many fielded items - with the surveys including both questionnaire and interview efforts. Proceeding through military channels, OFIG schedules five to eight trips a year to combat arms divisions that have just returned from major training exercises in the desert, jungle, or cold weather environments. The immediacy of the visits means that use of the items is fresh in mind, and the environmental harshness tests the level of abuse the items can withstand. Between 250 and 400 soldiers are surveyed on each occasion, and to date over 6,500 soldiers have been surveyed.

Project officers are the best source of information about a product and their input is the backbone of the questions asked in the surveys. This base is then fleshed out with questions necessary from a human factors perspective.

Besides the obvious payoff in specific product information, there are other rewards to systematic surveys and the data they produce. The data bases provide a frame of reference for product complaints. When a problem with a product is referred from the field, OFIG can quickly determine whether or not it is an isolated defect. Conversely, a few complaints can generate an item's being included on a future survey to determine if there is indeed a problem.

OFIG also uses the questionnaires and interviews to provide a clearer picture of the user. They are laced with questions about mission requirements, garrison and field life, and even extend to hygiene and eating habits. In short, OFIG asks anything that bears directly or indirectly on product design.

User Evaluations. After the survey process was fully operational, OFIG extended its sphere to user field tests of

developmental or modified items and in three years has evaluated 25 products/systems. While field testing is the ultimate test of a product, it does sacrifice scientific control. It sometimes also has to take an expedient path because of manpower and money constraints. However, OFIG operates on the principle that much can be gained from field evaluations in spite of their drawbacks.

We use a number of testing procedures. The simplest is to deliver a prototype item to a user for a predetermined time period. Usually there are only a few items available for testing, but at this stage the only goal is to identify any gross defects in design or function. Necessary modifications can then be made preparatory to a more comprehensive evaluation. Large amounts are not spent to produce several conceptual items, and testing is "piggybacked" onto a more comprehensive test or survey effort.

A specific instance is a recent evaluation of a prototype Combat Vehicle Crewman's Equipment Bag. OFIG delivered the one existing bag to members of an Armored Cavalry unit in Europe while on a survey trip and retrieved it 90 days later while on another evaluation. We interviewed the five personnel that had used it on a number of field exercises and sent their comments to the project officer. The soldiers felt it would make functioning in a tank easier, as it created more space by consolidating gear. Their one complaint was that it didn't have enough compartments; they wanted immediate access to some items when they had to "move out." Most of this information would not have surfaced from laboratory testing or from a project officer's intuition.

The other extreme of OFIG's field testing methods can be illustrated by a recent glove evaluation, which supported the Army's quest for a warmer glove for a moderately cold climate. It was conducted at three bases in the United States and one overseas and involved a total of 1400 soldiers who were assigned to either a control or experimental group. At the outset, the test soldiers were carefully fitted with the gloves and instructed on test protocol. At the end, data were collected on over 20 variables, and supportive weather and mission data were also obtained. This was relatively costly, but nothing compared to what the procurement of an inadequate product would exact in either dollars or dismay.

Most user evaluations fall at some point on the range bounded by the two extremes just described. The most common scenario is to have a product evaluated at one site, with approximately 30 to 60 users in both control and experimental groups. The duration of the tests is usually contingent on the time we feel we needed to assess any particular product's durability.

Other User Feedback. In addition to conducting surveys and evaluations OFIG attempts to help open communication lines and broaden the Army's knowledge of Natick's products.

For the most part, combat arms personnel have no clear picture of the structure and responsibilities of logistical support agencies such as Natick. OFIG has therefore made it policy to give users a formal, comprehensive briefing on Natick and a demonstration of its new and developmental products. Soldiers then know what products Natick is responsible for, what new ones they will be getting and when, and to whom to turn with problems, suggestions or needs. These briefings are given to personnel being surveyed, test personnel, and the command and staff of all divisions visited.

The feedback loop also extends to central issue facilities (CIFs), the organizations which maintain, store, and issue field equipment and clothing. OFIG calls on the manager at each site it visits to find out how Natick's products are performing, inspects the products, and advises on their proper care and issue.

OFIG also maintains a telephone hotline. It informs all users of its existence in the course of briefings, leaves cards and posters with the hotline number at military bases, and advertises it in military publications.

One last effort OFIG makes to close the developer-user gap is to include project officers in its visits to the military installations. A number of them have taken advantage of the opportunity and have returned with a better sense of what requirements their product has to meet and how the military community operates.

In-house Communications. Good communications are a key to any successful feedback operation and OFIG has established a wide and effective network at Natick. Its initial communications were devoted to informing all Natick personnel of its existence and its mission through briefings and memoranda. These communications are repeated and updated yearly for new project officers.

Reports are generated from every survey and user evaluation. Each is circulated not only to management and item project officers, but also to anyone that could conceivably benefit. That includes those responsible for similar products or those who have to assess factors such as their product's compatibility with the products in the report at hand.

Trip reports serve as the vehicle for communicating feedback obtained from commanders and staff of military units, CIF personnel, and other sources in the user world. Trip reports are also circulated to all who could possibly benefit from the information.

DISCUSSION

Overall Program. OFIG is not yet fully evolved. It has continued to grow in both personnel and responsibility during its existence. However, questions are starting to be raised about how much can be committed to a testing program in personnel or funding. While there is no doubt that testing must be conducted early to offset greater costs that would be incurred if design changes had to be made late in the development cycle, there are still decisions to be made as to the scope and frequency of such testing in order to be cost effective.

The user surveys also are leading to questions. How often does an item have to be surveyed to most effectively detect problems in manufacturing or basic design issues? What constitutes a solid, reliable data base for a product given the myriad of career fields that use it and the number of environments in which it is used? In other words, which cost benefit ratio in obtaining user feedback is the most advantageous? The attempts at answering these types of questions will most certainly be responsible for shaping OFIG's future.

Lessons Learned. The two greatest lessons learned so far are that the user unit's command support is critical to the success of the evaluation and that even under the best of circumstances, subject attrition is appalling. OFIG has learned to ask for at least twice the number of participants any test design would normally require. It also asks for hand receipts for the test products, puts requests to units for test support in writing, and directs these requests to a command level no lower than division.

In spite of these procedures, data acquisition can be problematic, and has had to be pursued relentlessly. In most cases the problem stems from the fact that there are many mission requirements in the field and upon returning from it. Access to the soldiers for purposes of collecting data becomes a lower priority than tanks that require maintenance or some other pressing mission demand. The most ideal scenario for collecting data would be while the soldiers are in the field, but this is often not permitted for logistical reasons, or because training schedules are often not compatible with data collection requirements. While we continue to search for field scenarios that accommodate our testing needs, we presently do the next best thing - survey immediately upon return from the field. We usually plan to spend extra days at a site if necessary so we can access as many subjects as possible, since many are often unavailable because of maintenance and other post-field requirements.

Another lesson OFIG learned in field testing was to keep test designs simple, as lack of control in the field is more than a figure of speech, and even simple designs can court testing disasters. A few examples of the types of situations that can

occur: (1) In conducting a test where soldier performance is an issue, one of the test companies can unexpectedly have to conduct night training, resulting in sleep deprivation on a day where performance was to be measured. (2) On a large scale field ration test one of the test groups could receive the wrong ration for the day. The supply point is a four-hour drive, the troops are moving out momentarily, and will be widely spread out for the rest of the day. (3) It rains all day on a day that troops were given questionnaires to fill out at a specific time point. In spite of waterproof bags being supplied the pervading moisture renders much of the data illegible.

When problems such as these occur when the design is simple, any complexity will increase the chances for disaster. It has been our experience that even simple crossover designs disintegrate. For instance, nature will assuredly supply a drastic change in the weather at a crossover point when testing different cold weather clothing items. Even in January there will be an unprecedented warm weather spell.

Crossovers may also not occur at the designated timepoint, because the test soldiers are not at the crossover location for unanticipated reasons - such as a unit being given the wrong map coordinates. Crossovers tend not to occur also when unit personnel are responsible for effecting the crossover, a situation sometimes agreed to because our evaluators are not always allowed access to some training sites. Despite the best intentions of unit personnel the exigencies of field life will always have priority, and people not familiar with research procedures do not recognize the importance of adhering strictly to test protocol.

The survey efforts are less of a problem in terms of support because they can be planned a year or more in advance. With a long lead time Natick can request access to users through a very high level troop command, and since these requests filter down through the combat arms chain, they have a higher priority.

CONCLUSION

At this point OFIG has a workable, productive program for obtaining user feedback. It has incorporated the user at all levels and in all combat arms branches. It has a number of vehicles for obtaining user information: surveys, product evaluations, interaction with unit commanders and their staffs, interaction with CIF's, and a hotline. It has effective in-house communication strategies.

Because this program is responding to the needs of a **unique** military multiproduct developer, it can hardly be considered a general model for obtaining user feedback. OFIG does, however, hope that its program will provide a frame of reference or at least a starting point for others who are looking for ways to involve the user in product feedback.

HOW MUCH DO SOLDIERS SLEEP?

Sally J. Van Nostrand

US Army Concepts Analysis Agency (CAA)

(Student research project, Industrial College of the Armed Forces, Class of '88)

Because of the dramatic effects to performance, sleep loss is one of the first variables that should be added to present combat models. However, when I began to plan the modifications to CAA models, I discovered a disconcerting lack of data on how much soldiers sleep.

METHOD

I administered a questionnaire to officers and noncommissioned officers (NCOs) at CAA and to Army students and faculty at the National Defense University. Nearly all of the subjects have held command positions, served on the staffs at several levels, and have participated in exercises such as Reforger in Europe and at the National Training Center (NTC). Many took the trouble to write additional comments; one of the most thoughtful sets of comments was provided by a senior NCO who had a wide range of experience.

The questionnaire was relatively simple--for various generic duty positions, I asked the respondents to tell me the number of hours they expected to work, sleep, and perform other activities during combat conditions, and to circle the item numbers of the duty positions which they had actually held. They completed items for only those positions they had held or for which they felt they had personal knowledge.

RESULTS

From the 95 questionnaires returned, I obtained 752 estimates of the average amount of sleep expected under combat conditions. The duties covered by the estimates range from cook to division commander. The largest numbers of duty position estimates were for battalion staff, company commander, and platoon leader (83, 84, and 78, respectively); nearly all of these estimates were from officers who had served in those positions, plus a few for battalion staff from NCOs who had served on one. Since most of the respondents were officers and they only provided estimates for duties they felt comfortable about estimating, there are many fewer estimates for each of the enlisted and NCO positions.

Figure 1 shows groupings of an assortment of the duties by the average number of hours of sleep estimated. This graph clearly shows a large range in the amount of sleep expected by duty position. The combat officer and combat NCO seem to expect the least amounts of sleep (4.5 hours and 4.4 hours, respectively). In all of these figures you will notice that aviation soldiers are shown separately, and that they have some of the highest expected averages. It seemed reasonable to separate them because aviation soldiers in flying status are required to sleep at least 6 hours in one unbroken segment. Another interesting aspect of the aviation officer data is that the least amount of sleep they expect is 3 hours. There

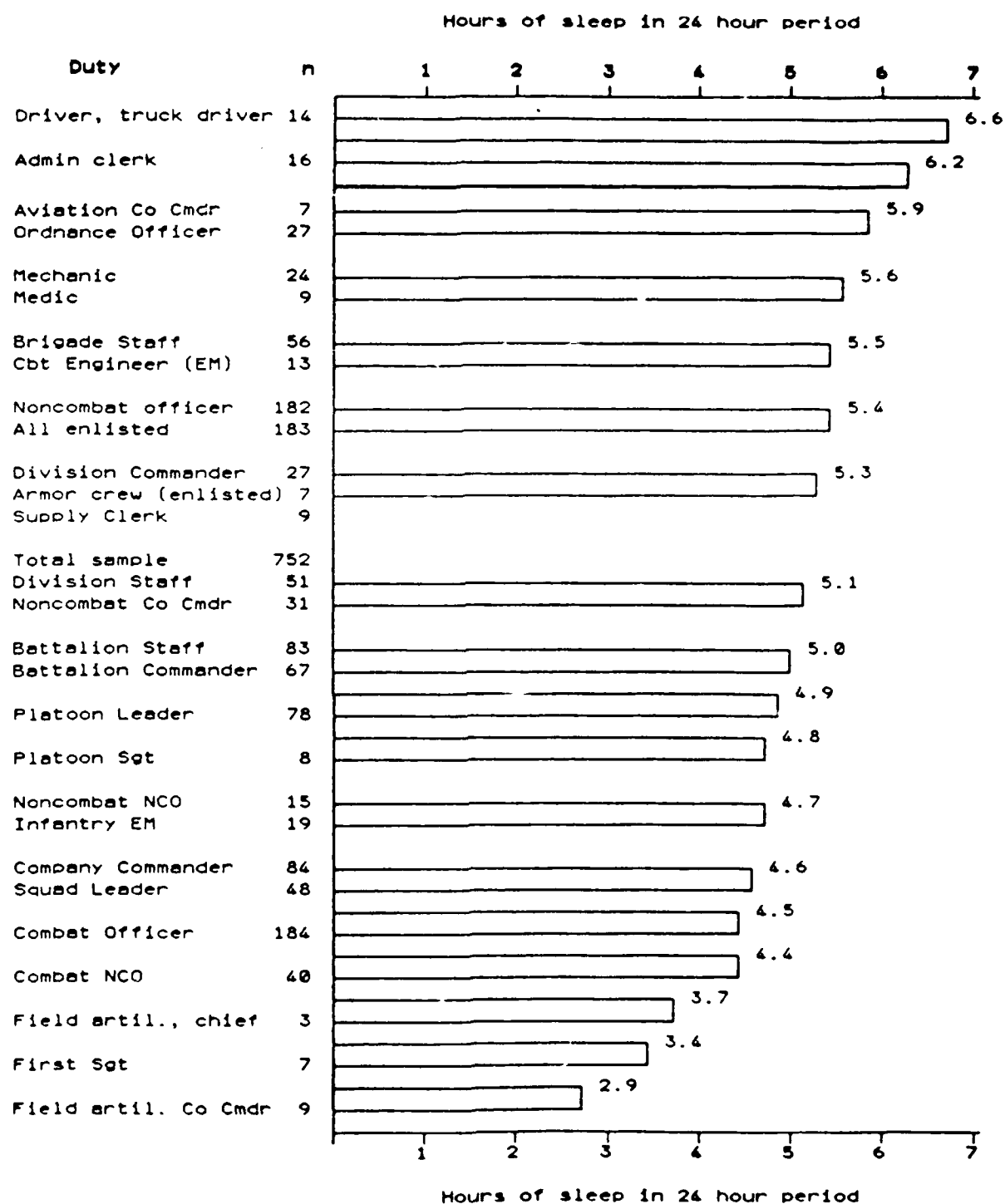


Figure 1. Selected duty positions--grouped by expected number of hours of sleep in a 24 hour period

were several in each of the other categories that believed they would sometimes have to go entirely without sleep.

These data show that most US Army soldiers should expect some sleep decrement during combat conditions (see Figure 2). Officers expect that the average noncombat enlisted soldier will get nearly the required 6 hours (some will sleep more) and nearly all enlisted in aviation-related jobs will sleep more than 6 hours. However, NCOs, even in aviation-related positions are predicted to have substantial sleep decrements.

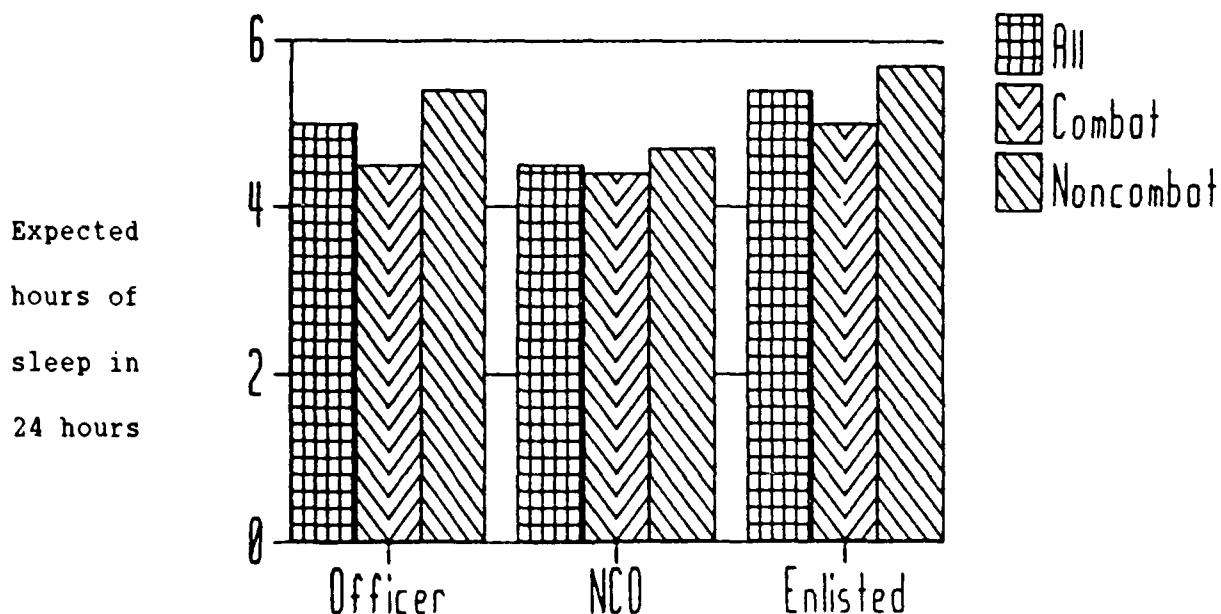


Figure 2. Officer, NCO and enlisted average expected amounts of sleep

Figure 3 shows officer sleep loss by duty position. There is a large difference between the amount of sleep that aviation officers expect and the amount that other officers can expect, particularly the combat platoon leaders and combat company commanders. Combat company commanders are estimated at less than two-thirds of the minimum required amount.

Within the company commander position there are also large differences. As Figure 4 shows, the company commanders of units that are in direct contact with the enemy are expected to sleep less than 4 hours out of each 24 hours (Infantry and Armor). Field artillery company commanders expect to sleep less than half of the needed 6 hours (2.9 hours). Although the noncombat company commanders are sleeping more than the combat company commanders, only the aviation company commander comes within a half hour of the required number of hours. Although the effect of the noncombat commanders' decisions may not have the immediacy of those of the combat commanders, they must affect the total war effort.

Expected
hours of
sleep in
24 hours

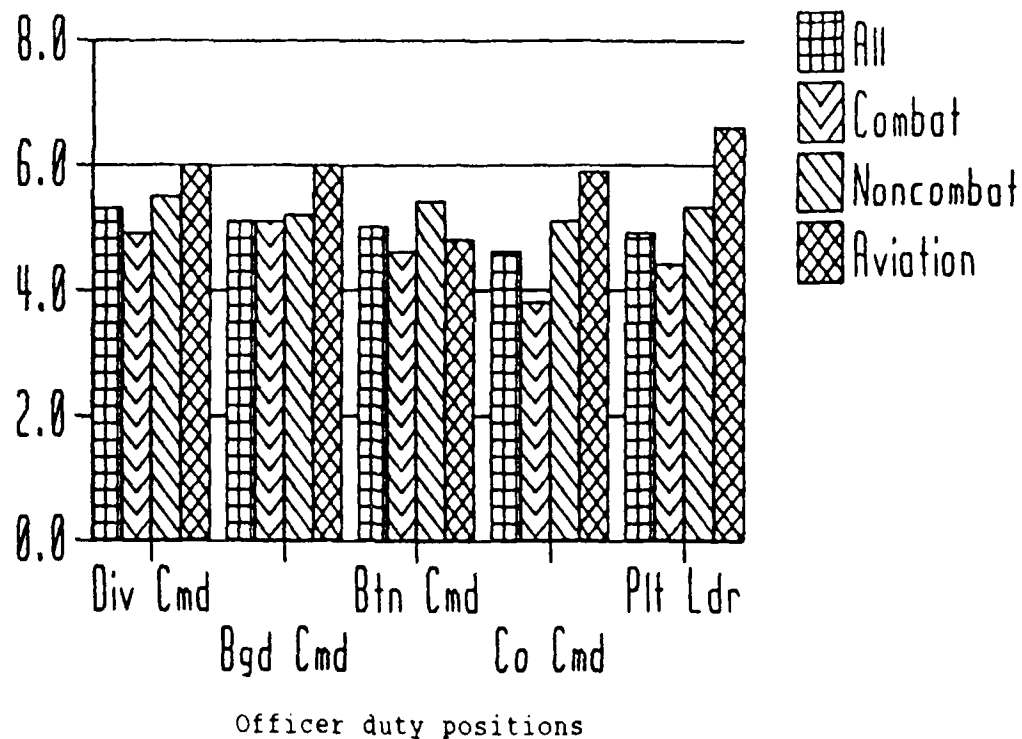


Figure 3. Expected officer sleep

Expected
hours of
sleep in
24 hours

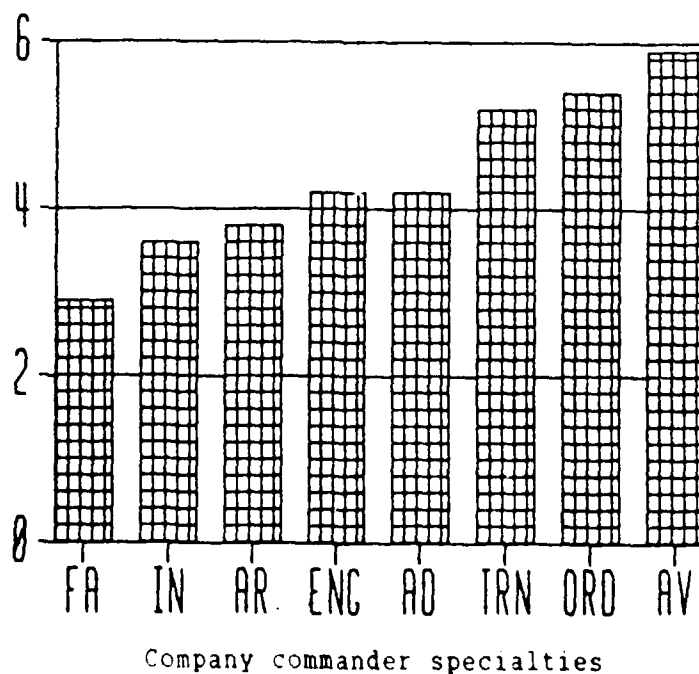


Figure 4. Company commander expected sleep

DISCUSSION

Are soldiers who sleep less than the desired 6 to 8 hours functioning at 100 percent of their capability? Research shows that the average person is functioning on a sleep decrement with less than 6 to 8 hours. Although they can function physically and seem able to make most decisions, they are building a sleep debt. Eventually, this sleep debt builds to a point that the soldier will break. How big this sleep debt must be to cause a break, and how much it varies from person to person, is as yet unknown, but it does exist.

From conversations with military personnel, particularly in the combat specialties, and from reading accounts of combat, I believe that most military personnel think they can function on as little as 4 hours. They know they feel better with more, and they would like to have more. However, since they are able to stay awake with 4 hours, they believe they are functioning well. This (from an aviation officer) is one of several similar comments: "It was occasionally necessary to work 2 or 3 days straight. We couldn't keep it up indefinitely, but needed to do it once in a while in every scenario."

The soldiers who play the Red side at NTC (opposition force or OPFOR) are taught the necessity of sleep. According to one respondent, the NTC doctor works with them on both their "conditioning and sleep plan--they go together." He also had this to say about the units who rotate through NTC for training: "My 30 months [in different positions in the NTC OPFOR] taught me that leaders must sleep. I've seen many go for 2-4 days without sleep, but always, they paid for it by not making timely or sound tactical decisions--their units died. The OPFOR had a sleep plan. No matter how bad things got leaders got at least 6 hours sleep in one shot. We were in the field over 200 days a year."

It is interesting that some segments of the military (aviation and NTC OPFOR) know that 6 hours, or more, in one unbroken segment is necessary if people are not to make dumb decisions, and yet this knowledge is not passed on to the rest of the military community. Even soldiers who should know it do not generalize the need to make smart decisions while performing other duties. The aviation officer above should have known better, yet he estimated average amounts of sleep as small as 4 hours except when in flying status.

Even though many officers seem to know that a sleep plan is necessary they do not seem to know what it should be. One comment (includes his original emphasis) was, "In combat or combat-like conditions, all soldiers work extremely long hours under very arduous conditions. It is imperative that commanders of large units establish a good sleep plan. If not, after 3 days his unit becomes ineffective!" Yet this officer used estimates as small as 4 hours of sleep for some duty positions.

Can we quantify the relationship of sleep decrements and soldier performance?

Given that we agree that less than 6 hours (used by aviation and NTC OPFOR) is a sleep decrement, then what does this sleep decrement mean? A decrement of "only a couple of hours" may not seem like very much. However, 2 hours is one third of the required 6 hours. I have not been able to determine how much we should decrement a soldier's performance other than for no sleep in a 24 hour period--then the performance decrement is 25 percent. If we can assume that the decrement is the same for each hour of sleep lost, and if 6 hours is a reasonable number to use for the beginning requirement, then the decrement is 4.17% per hour

of sleep lost. At this rate, the average combat officer takes 13 days to reach less than 50 percent of capacity performance. Figure 5 shows this theoretical curve for combat officers, combat company commanders and Field Artillery company commanders. Pretty bleak, isn't it?

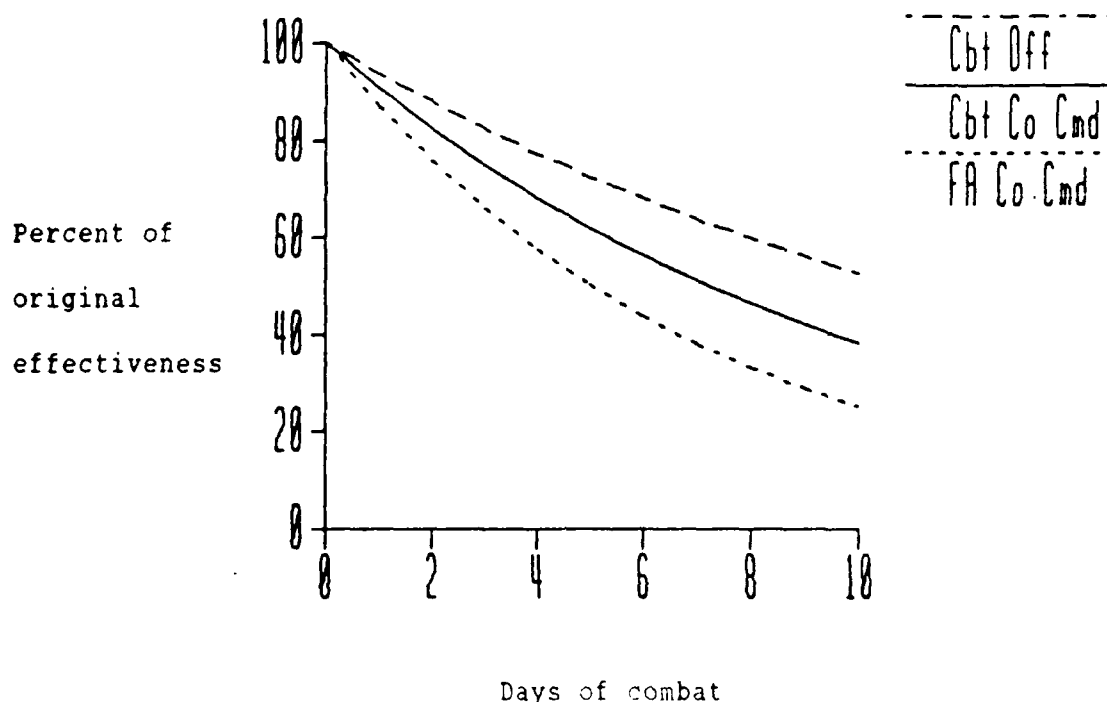


Figure 5. Theoretical effectiveness as a function of sleep loss

SUMMARY OF QUANTIFICATION OF SLEEP LOSS ON THE BATTLEFIELD

The results showed that only aviation soldiers and some enlisted specialties should expect to sleep the requisite 6 hours in each 24 hour period; all other soldiers will build a sleep debt. Combat soldiers will sleep the least. Company commanders will sleep only one-half to two-thirds the required amount, and very few combat soldiers can expect more than 5 hours. Noncombat officers and NCOs should also expect to build a sleep debt--noncombat NCOs will sleep very little more than will the combat NCOs. These results show that units should not be modeled at 100 percent effectiveness. From sleep loss only, combat units will probably lose 6.25 percent of their effectiveness each day they are in contact with the enemy. Noncombat units will also lose effectiveness, but not at such a high rate as the combat units, perhaps as much as 4 percent per day. During the course of a protracted war, noncombat units, because they are never relieved, may degrade more than the average combat unit. These results, also, show that commanders do not know some fundamental aspects of human behavior and sleep requirements.

TESTING VISION, PERFORMANCE, AND SUBJECTIVE SYMPTOMATOLOGY
IN REDUCED OXYGEN ENVIRONMENTS

Richard L. Burse, Sc.D., Chair
U.S. Army Research Institute of Environmental Medicine
Natick, Massachusetts

Reducing the concentration of oxygen (O_2) in air below the normal value of 21% through nitrogen dilution has been proposed as a means of increasing fire safety in confined spaces. Reducing the O_2 to 19% inhibits combustion somewhat without noticeably affecting occupants, but the O_2 concentrations at which perception, performance, and affect are altered have yet to be identified. Accordingly, four related experiments were undertaken in a joint study sponsored by the US Navy in collaboration with the US Army to discover the impact of 3-day exposures to 21%, 17%, and 13% O_2 environments on vision, cognitive and psychomotor performance, mood, and altitude illness symptomatology.

Three principal conclusions appear warranted by the presented results of the four panelists. First, and most important, 17% appears to have had no determinable impact on scotopic visual sensitivity, cognitive and psychomotor performance tasks, mood, or altitude symptomatology tested during the study. Secondly, 13% oxygen, and 17% oxygen under reduced pressure which brought the PO_2 down to that of 13% O_2 , had no effect on scotopic visual sensitivity and no or only transient effect on cognition/motor performance, but did have effects on mood and altitude illness symptomatology similar to those previously seen in mountaineers at the same PO_2 . Thirdly, none of the cognitive/psychomotor tasks showed stable performance at the end of the study, even after an 8-day training and 15-day study period. This suggests the need for a final control period following the last experimental period in order to correctly determine the extent of the training effect still operating during the last experimental period.

VISUAL SENSITIVITIES UNDER REDUCED OXYGEN
S.M. Luria and Nancy Morris
Naval Submarine Medical Research Laboratory

and Alan Cymerman
U.S. Army Institute of Environmental Medicine

Many investigators have proposed that visual sensitivity may be the most sensitive measure of oxygen deprivation (Halperin, et al., 1947; Carr, et al., 1966; Crews, 1966). Mountain climbers have commented on the noticeable dimming of vision with hypoxia (Hornbein, 1983; Griffith, et al., 1983).

There has, however, been considerable disagreement as to the level of oxygen deprivation at which visual thresholds are affected. McFarland (1970) and his colleagues have reported degradations of visual thresholds at altitudes of less than 5,000 feet, when arterial oxygen saturation was decreased less than 5%. Indeed, Halperin et al. (1959) reported measurable changes in visual threshold during an experiment session while carbon-monoxide levels were being changed.

On the other hand, several investigators have found no significant changes in visual measures at carboxyhemoglobin levels of as much as 20%, corresponding to an altitude of 15,000 feet (Beard and Grandstaff, 1970; Stewart et al., 1970; Weir et al., 1973; Roche et al., 1981).

In our previous studies, we have not found decrements in the scotopic sensitivity of subjects exposed for three hours to 200 ppm of carbon monoxide (Luria and Schlichting, 1979) or of subjects breathing air containing only 13% oxygen for three hours (Knight et al., 1987). Rather, we have found scotopic sensitivity to be affected only after the oxygen level in the air was reduced to 10% (Luria and Knight, 1987).

Similarly, there has been a lack of agreement on the effect of anoxia on the limits of the visual field (Tune, 1964). For example, Kobrick showed in several studies (see Kobrick, 1976) that the detection of peripheral stimuli was degraded as altitude increased. But Birren et al. (1946) found no changes even at 18,000 feet where the partial pressure of oxygen is half that at sea-level.

In this study, we measured the effects of 13 and 17% oxygen on night vision sensitivity and field of view using a double-blind protocol.

METHOD

Subjects. Thirteen active duty Navy and Marine Corps volunteers (aged 18 to 30 with a mean of 24 years) completed the experiment.

General Procedure. After two practice sessions a day for a week at sea-level pressure, the subjects were sealed in the

hypobaric chamber for 15 days. The oxygen level was 21% for the first three days, 17% for the next three days, 21% for the next three days, 13% for the next three days, and 21% for the final three days. In addition, during the last seven hours of the 17% condition, the barometric pressure was reduced to 576 torr to simulate the partial vacuum caused by a "high vacuum diesel shutdown". The oxygen levels were changed without the knowledge of either the subjects or the experimenter. Under each oxygen condition, they were tested twice a day on the first day, on the afternoon of the second day, and twice on the third day.

Scotopic Sensitivity. The subjects were dark-adapted for 20 minutes and then led into a light-proof chamber. They occluded the left eye with an eye-patch, put on a headset to communicate with the experimenter, and positioned their heads in a chin-rest. This placed their eye 60 cm from a ground-glass screen on which was projected a circle of light subtending 0.5 deg visual angle. This was 10 deg to the left of a red pin-point fixation light and was flickered at 2 cps to facilitate recognition. The dominant wavelength of the light was 500 nm; its intensity was varied with neutral density filters. Thresholds were measured with the Method of Constant Stimuli and were taken as the 50% point.

Perimetry. A semi-circular band of aluminum, 70 cm in diameter and about 3 cm wide, was worn on the head (Fig. 1). There were six sets of light-emitting diodes (LEDs) on each side with a red and green LED in each set. The first set made an angle of 54 deg with the line of sight directly to the front, and the last set made an angle of 96 deg. The visual fields were measured under ordinary room illumination. The subject fixated a white dot on the band directly in front of him. The red stimuli were presented first. At irregular intervals, an LED would be flashed briefly, and the subject responded "right" or "left" if he saw it. There were 5 to 10 trials for each test light, depending on the subject's variability. The procedure was repeated for the green stimuli. The calculated angles at which each color was seen 50% of the time on each side were averaged, and this was taken as the threshold.

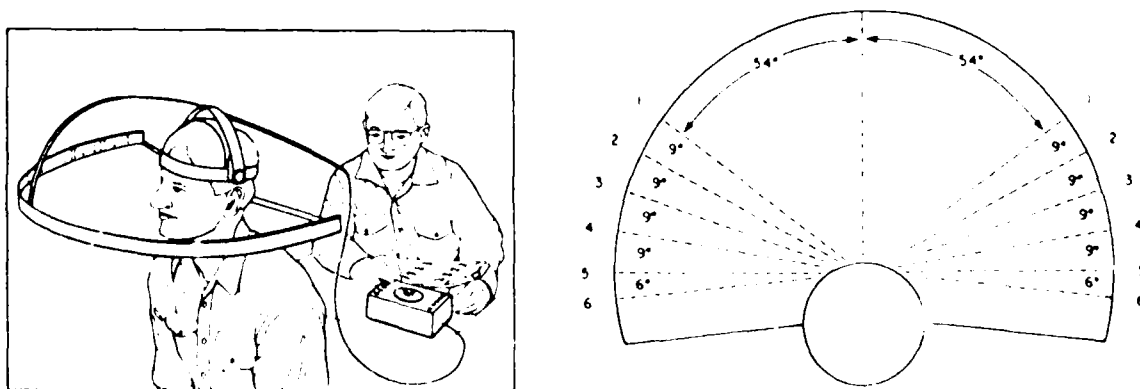


Figure 1. The perimeter

RESULTS

Scotopic Sensitivity. Figure 2 shows the mean scotopic thresholds for each condition. There is no indication of a drop in sensitivity when the oxygen level was lowered. The final threshold in the 17% condition was obtained under the partial vacuum. Neither that specific threshold nor the mean threshold for the condition was unduly poor compared with the other thresholds. There were no significant differences between the conditions according to a one-way repeated measures analysis of variance.

We also compared the first thresholds obtained during exposure to each condition-- that is, after about three hours of exposure-- with the final thresholds in each condition-- that is, after about 52 hours of exposure. Figure 3 shows no indication that either of the reduced oxygen conditions degraded sensitivity.

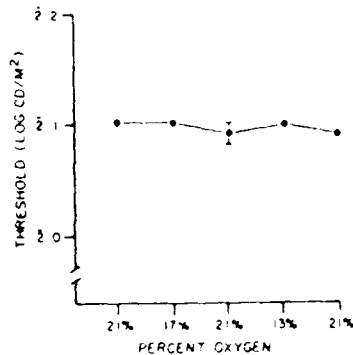


Figure 2. Mean scotopic thresholds for each oxygen level.

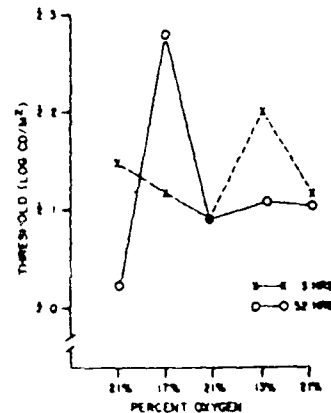


Figure 3. Mean scotopic thresholds at the start and end of each oxygen condition.

Field of View. Figure 4 shows the mean perimetry thresholds for each color in each condition for the seven subjects who completed this procedure. The threshold improved significantly during the course of the study ($F(4,24) = 6.75$, $p < .01$) according to an analysis of variance. This appears to be a practice effect despite the initial week of practice. But there were no apparent degradations in the field of view when the oxygen level was reduced; there were no significant differences between the low oxygen conditions and the adjacent 21% conditions. And there were no differences between the colors. Finally, there were again no apparent differences between the thresholds taken at the beginning of each condition and those taken at the end of the condition (Figure 5).

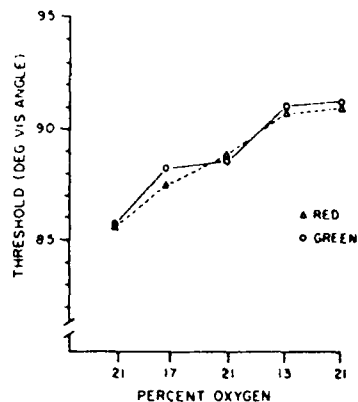


Figure 4. Mean perimetry thresholds for each oxygen condition.

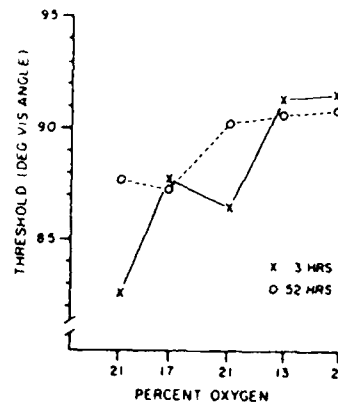


Figure 5. Mean perimetry thresholds at the start and end of each oxygen condition.

DISCUSSION

There was no indication in these results of any deleterious effects of hypoxia on either scotopic sensitivity or field of view. We conclude that reducing the percentage of oxygen in the breathing mixture to 13% for three days does not degrade these visual processes. As in our previous studies (cf. Luria and Knight, 1987), the results do not support the conclusion that vision is especially sensitive to small reduction in the level of oxygen. Rather, they support the findings that rather sizable reductions in oxygen are required to degrade vision (Gellhorn, 1936; Otis, 1946; Tune, 1964; Pierson, 1967; Fowler, 1985). We have found this result repeatedly in double-blind experiments, and we conclude that this design is essential in these experiments.

DISCLAIMER

Naval Medical Research and Development Command, Navy Department, Research Work Unit No. 61152N MR00001.01-5103. The views in this article are those of the authors and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.

REFERENCES

- Beard, R.R. and Grandstaff, N. (1970). Carbon monoxide exposure and cerebral function. In: Coburn, R.F. (Ed.) *Biological Effects of Carbon Monoxide*. Ann. N.Y. Acad. Sci. 174: 385-395.

- Birren, J.E., Fisher, M.B., Volmer, E., and King, B.G. (1946). Effects of anoxia on performance at several simulated altitudes. J. Exp. Psych. 36: 35-49.
- Carr, R.E., Gouras, P., and Gunkel, R.D. (1966). Chloroquine retinopathy: early detection by retinal threshold test. Arch. Ophthalmol. 75: 171-178.
- Crews, S.J. (1966). The prevention of drug induced retinopathies. Trans. Ophthalmol. Soc. UK 36: 63-76.
- Fowler, B., Paul, M., Porlier, G., Elcombe, D.D., and Taylor, M. (1985). A re-evaluation of the minimum altitude at which hypoxic performance decrements can be detected. Ergonomics 28: 781-791.
- Gellhorn, E. (1936). The effects of O_2 -lack, variations in the CO_2 -content of the inspired air, and hyperpnea on visual intensity discrimination. Am. J. Physiol. 115: 679-684.
- Griffith, L., Pugh, C.E., and Sutton, J.R. (1983). Everest then and now. In: J.R. Sutton, C.S. Houston, and N.L. Jones (Eds.), Hypoxia, Exercise, and Altitude, New York: Alan R. Liss, Inc., pp. 415-428.
- Halperin, M.H., Niven, J.I., McFarland, R.A., and Roughton, F.J.W. (1947). Variations in visual thresholds during carbon monoxide and hypoxic anoxia. Fed. Proc. (Abs.) 6: 120-121.
- Halperin, M.H., McFarland, R.A., Niven, J.I., and F.J.W. Roughton (1959). The time course of the effects of carbon monoxide on visual thresholds. J. Physiol. 146: 583-593.
- Hornbein, T.F. (1983). Everest without oxygen. In: J.R. Sutton, C.S. Houston, and N.L. Jones (Eds.), Hypoxia, Exercise, and Altitude, New York: Alan R. Liss, Inc., pp. 409-414.
- Knight, D.R., Luria, S.M., Socks, J.F., and Rogers, W. (1987). Effect of nitrogen-based, fire-retardant atmospheres on visual and mental performance. In: Bove, A., Bachrach, A., and Greenbaum, L. (Eds.), Underwater and Hyperbaric Physiology IX. Proc. Ninth International Symposium on Underwater and Hyperbaric Physiology. Bethesda, MD: Undersea and Hyperbaric Medical Soc., pp. 521-534.
- Kobrick, J.L. (1976). Effects of prior hypoxia exposure on visual target detection during later more severe hypoxia. Percept. Mot. Skills 42: 751-761.
- Luria, S.M. and Knight, D.R. (1987). Scotopic sensitivity with 10% oxygen. NSMRL Rep. No. 1097. Groton, CT: Naval Submarine Medical Research Laboratory.
- Luria, S.M. and Schlenting, C.L. (1979). Effects of low levels of carbon monoxide on vision of smokers and nonsmokers. Arch. Envir. Health 34: 38-44.

- Otis, A.B., Rahn, H., Epstein, M.A., and Fenn, W.O. (1946). Performance as related to composition of alveolar air. Am. J. Physiol. 146: 247-253.
- Pierson, W.R. (1967). Night vision and mild hypoxia. Aerosp. Med. 38: 993-994.
- Roche, S., Horvath, S.M., Gliner, J.A., Wagner, J.A., and Borgia, J. (1981). Sustained visual attention and carbon monoxide: Elimination of adaptation effects. Hum. Fact. 23: 175-184.
- Stewart, R.D., Peterson, J.F., Baretta, E.D., Bachland, R.T., Hosko, M.J., and Hermann, A.A. (1970). Experimental human exposure to carbon monoxide. Arch. Environ. Health 27: 155-160.
- Tune, G.S. (1964). Psychological effects of hypoxia: review of certain literature from the period 1950 to 1963. Percept. Mot. Skills 19: 551-562.
- Weir, F.W., Rockwell, T.H., Mehta, M.M., et al. (1973). An Investigation of the Effects of CO in Humans on the Driving Task. Columbus, Ohio: Ohio State University Research Foundation.

COGNITIVE AND MOTOR PERFORMANCE UNDER REDUCED OXYGEN

by

Christine L. Schlichting, Douglas R. Knight
Naval Submarine Medical Research Laboratory

and Alan Cymerman
United States Army Research Institute of Environmental Medicine

There continues to be some controversy over how low the percentage of oxygen in a breathing mixture can be reduced before the hypoxia affects performance. Historically this question has been of interest to aviators and to astronomers in high altitude observatories. More recently the question has again been raised as the Naval community studies possible ways to reduce the hazard of fires on submarines. There is a substantial literature in this area but unfortunately there is no clear answer to the performance question.

In 1964, G. S. Tune suggested that an altitude of 10,000 feet might be an upper limit beyond which significant effects occur. This altitude is equivalent to breathing 14.4% oxygen at sea level. Later reviews by Ernsting (1978, 1984) concluded that the hypoxia equivalent of breathing air at 5,000 feet (17.4% oxygen) represented an acceptable compromise for performance in aircraft crews.

Other authors feel that these levels are too conservative and that acceptable performance is possible at higher altitudes - lower SaO₂s (arterial oxyhemoglobin saturation). Fowler, Paul, Porlier, Elcombe and Taylor (1985) concluded that performance on a spatial transformation test was not affected at an SaO₂ of 88-90%, the altitude equivalent to 7,899 feet. Fowler, Elcombe, Kelso and Porlier (1987) used a serial choice reaction time task and concluded that response time was slowed in a "dose dependent" manner reaching a significant decrement at 82% SaO₂ which they calculate to be equivalent to 10,000 feet using a formula described in their paper.

Although these studies are suggestive of hypoxia thresholds, some questions still remain. It would be premature to attempt to generalize from only two tasks to the many jobs required in a military setting. Moreover, these studies do not address adaptation effects. Finally, effects on motor performance alone have not been studied at intermediate levels of hypoxia.

The study reported here was designed to measure hypoxic effects on motor and cognitive performance at the hypoxic levels (equivalent altitude of 5,000 to 12,600 feet) of interest to the submarine community. Since slow ascents to altitude confer protection against hypoxia to mountaineers, we also included a staged reduction of chamber pO₂ to 99 torr as one of the conditions. This condition will determine the effect of adaptation to an intermediate level of hypoxia (17% oxygen at sea level) on performance at a greater level of hypoxia (17% oxygen at hypobaric pressure).

Method

Subjects

Thirteen enlisted men from the U.S. Navy and Marine Corps completed all of the phases of the study. They ranged in age from 18-36 years with a mean age of 24 years. They gave informed consent and were medically screened.

Tests

Tests were chosen that met the criteria for repeated measures testing as defined by Bittner, Carter, and Kennedy (1977). The test battery included paced mental arithmetic (vertical addition), 4-choice reaction time, code substitution, and Grammatical Reasoning. The four tests of motor skill used were the Minnesota Rate of Manipulation, the O'Connor Finger Dexterity, Rotary Pursuit and Precision-Steadiness. Order of performance of the tests was random within subjects across test sessions. Tests were either self-administered or administered by another subject.

General Procedure

Each subject practiced each test twice a day for a week in the USARIEM hypobaric chamber. Subjects were then sealed in the chamber for 15 days. The oxygen level was maintained at 21% for 72 hours then changed to 17% ($PO_2 = 129$ torr) for 56 hours, then 17% at reduced barometric pressure ($PO_2 = 98$) for 7 hours, 21% for 81 hours, 13% ($PO_2 = 99$ torr) for 63 hours, and back to 21% for the final 72 hours. Carbon dioxide was added to each breathing mixture (.9%). Barometric pressure ranged from 744 to 771 torr except on day three when the pressure was reduced to 576 torr to produce a PO_2 torr of 98.

In each condition, subjects were tested twice on the first day, once on the second day and twice on the third day. Because of the restricted size of the hypobaric chamber, the study was run in two identical phases with six subjects in the first and seven in the second phase of the experiment. Some data is missing for the 13% exposure due to illness and noncompliance.

RESULTS

Mean SAO_2 values for the air, 17% oxygen and 13% oxygen conditions at rest were 98.4 (S.D. .9) at 20.9%, 97.2 (S.D. .7) at 17% and 92.0 (S.D. 2.1) at 13%.
Cognitive Tests

The grammatical reasoning test and the choice reaction time measures (number correct and the mean reaction time) showed no significant differences over the course of the experiment.

The number correct on the mental math showed a progressive increase in performance over the course of the study ($F=14.92$; $df=4,48$; $p<.01$) while the mean time per problem showed a decrease ($F=4.87$; $df=4,48$; $p<.01$). The Digit Symbol Substitution test also showed a gradual improvement in performance throughout the course of the experiment ($F=5.29$; $df=4,48$; $p<.01$).

Motor Tests

Some of the motor tests showed decrements in performance during the reduced oxygen conditions. There were no differences in performance across the five test sessions within each oxygen condition (lapse), nor was there any interaction of lapse by oxygen condition. Separate ANOVAs comparing performance during the 13%, 17% normobaric and 17% hypobaric conditions showed an effect only for the rotary pursuit test.

Minnesota Rate of Manipulation. When the data were analyzed across sessions within a condition the Minnesota Rate of Manipulation ($F=19.16$; $df=4,48$; $p<.01$), showed a gradual improvement in performance.

Rotary Pursuit. The rotary pursuit ($F=12.76$; $df=4,48$; $p<.01$), showed significant differences across conditions with the worst performance at 13% and the best at the final 21% condition. Figure 1 shows the rotary pursuit data for

the first session of each of the five conditions and the mean of each condition.

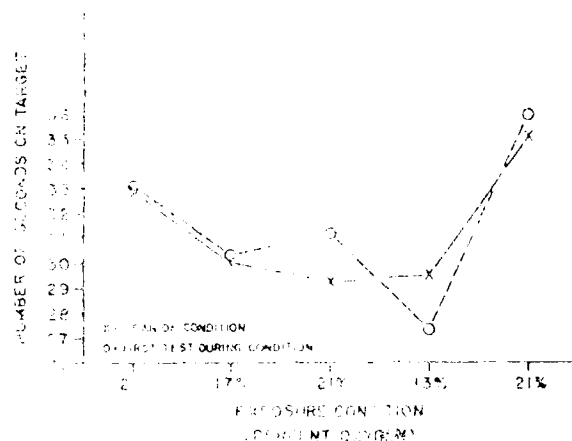


Figure 1. Rotary Pursuit Performance during different exposures.

Seven subjects took the rotary pursuit task during the first test session under 13% oxygen. When the data for only the first test session of each condition were analyzed, there was a decline in performance upon exposure to the 13% condition ($F=16.62$; $df=4,24$; $p<.01$) for the seven subjects that took the test. Performance during the hypobaric pressure condition at 17% was reduced from performance during the normobaric pressure exposure at 17% condition. Subjects averaged 32.17 seconds on target for the hypobaric condition versus 36.42 seconds for normobaric condition. This difference was significant at the .01 level on the Newman Keuls test, as was the difference between the 13% condition (average seconds = 30.96) and the normobaric condition at 17%. There was no significant difference between the 13% condition and the hypobaric condition.

Steadiness. The ANOVA showed that there was a significant difference among conditions for the steadiness test ($p<.01$). The data suggests gradual improvement in performance with decreased steadiness during the 13% condition. None of the post hoc comparisons reached significance.

When only the data for first test session were analyzed there was a significant difference in performance across the oxygen conditions ($F=7.23$; $df=4,18$; $p<.01$). All of the individual comparisons were significant on the Newman Keuls test. The data are plotted in Figure 2.



Figure 2. Steadiness and Finger Dexterity

O'Connor Finger Dexterity Test. There was a significant improvement in performance across conditions for the number of trials filled ($F=15.11$; $df=4,48$; $p<.01$). When the data for the first session of each condition were analyzed separately (Figure 3), the worst performance occurred during the 17% condition ($F=9.23$; $df=4,42$; $p<.01$). The Newman Kuuls test showed that this value was significantly lower than all subsequent values whereas values for the following three conditions did not differ among themselves. There are significant differences among conditions ($F = 9.89$, $df = 4,24$, $p < .01$) in this smaller sample, but none of the individual differences were significant.

DISCUSSION

The only tests to suggest any effect of reduced oxygen were the rotary pursuit, O'Connor finger dexterity test, and the Steadiness test. The O'Connor test showed a transient effect at the start of the 17% condition. Performance appeared to return to normal for the rest of that condition. Because of the reduced number of subjects, it would be premature to draw conclusions from the data collected during the 13% condition for this test. However, since half of the subjects chose not to complete the tests under these conditions, we suspect that performance for the first test session under 13% oxygen may also have been initially degraded.

The existence of a fine motor decrement at slightly higher levels of hypoxia has been reported. Portian (1972) found that performance on a finger dexterity task was initially degraded at 18,000 feet, but only steadiness was affected in altitude acclimatized subjects (6 to 15 days at altitude).

Data from both the rotary pursuit and the steadiness test suggest decreased motor performance during the 13% condition. Performance on the rotary improved during the rest of the condition but performance on the steadiness test remained depressed. Performance during the 17% hypobaric condition was not different from performance during the 13% condition on the rotary pursuit. This suggests that under these experimental conditions a staged decrease in pO_2 showed the same decrement in performance as a direct decrease in pO_2 .

A search of the literature did not reveal any investigations under hypoxic conditions that used a measure of motor finger dexterity or the Steadiness test at intermediate levels. In addition, although there are numerous reaction time and tracking studies (including those conducted by Fleishman and Ellison (1952) later on as a part of this study) and steadiness skills. This study suggests that the rotary pursuit and steadiness tests may be more sensitive to reduced oxygen than the reaction time and tracking tests. Further testing with large sample groups is needed to clarify the extent of motor performance decrements under hypoxia.

REFERENCES

- Portian, R. (1972). The effects of hypoxia on human performance. *Aviation, Space, and Environmental Medicine*, 43, 100-105.
- Performance on the majority of cognitive tests was extremely resistant to reduced oxygen at the hypobaric condition. Several other cognitive tests similarly showed no effect of hypoxia (e.g., Bower, 1978; Bower, Banderet, Knight, and Amerman, 1980; Bower, Banderet, and scotopic vision have also been reported to be unaffected by hypoxia (Kobrick, Grodin, Sherman, and Bower, 1980; Bower, Banderet, and Morris, 1988).
- Other studies have reported no effect of hypoxia on reaction time to hypoxia at the hypobaric condition (e.g., Bower, Banderet, and Winneham,

1968; Alnutt, 1968), reaction time, problem solving (Peterson and Neal, 1970) and card sorting (Kelman, Crow, and Bursill, 1969). Only a vigilance task (signal detection) and a complex fire control task (Fine and Kobrick, 1978) show performance decrements with this degree of hypoxia. At altitudes of 12,000 feet and above, investigators more consistently report performance decrements.

This study suggests that there will be little or no decrement in cognitive functions with levels of hypoxia approaching 17% (8,000 feet). Any transient effects of hypoxia on fine motor control will dissipate rapidly. We do not know if exposures considerably longer than three days will eventually have any deleterious effects. At the equivalent of 12,000 feet there may be more prolonged disturbances of fine motor control but cognitive functions appear to be unaffected.

References

- Allnutt, M. F. (1968) Complex mental function under heat and hypoxia. Aerospace Medical Congress.
- Bittner, A. C., Carter, E. C., & Kennedy, R. S. (1986) Performance evaluation tests for environmental research (PETER): Evaluation of 114 measures. Perceptual and Motor Skills, 63, 683-708.
- Carver, R. P., & Wimsman, F. R. (1968) Effect of high elevation upon physical proficiency, cognitive functioning and subjective symptomatology. Perceptual and Motor Skills, 26, 223-230.
- Ernsting, J. (1978) Prevention of hypoxia - Acceptable compromises. Aviation, Space and Environmental Medicine, 49, 495-502.
- Ernsting, J. (1984) Mild hypoxia and the use of oxygen in flight. Aviation, Space and Environmental Medicine, 55, 407-410.
- Fine, B. J., & Kobrick, J. L. (1978) Effects of altitude and heat on complex cognitive tasks. Human Factors, 20(1), 115-122.
- Fleishman, E. A., & Ellison, G. D. (1962) A Factor analysis of fine manipulative tests. Journal of Applied Psychology, 46, 96-105.
- Fowler, B., Elcombe, D. D., Kelso, B., & Porlier, G. (1987) The threshold for hypoxia effects on perceptual-motor performance. Human Factors, 29(1), 61-66.
- Fowler, B., Paul, M., Porlier, G., Elcombe, D. D., & Taylor, M. (1985) A re-evaluation of the minimum altitude at which hypoxic performance decrements can be detected. Ergonomics, 28, 781-791.
- Kelman, G. R., Crow, F. J., & Bursill, A. E. (1969) Effect of mild hypoxia on mental performance assessed by a test of selective attention. Aerospace Medicine, March, 261-263.
- Kobrick, J. L., Grohn, E., Shukitt, B., Houston, G. S., Sutton, J. R., & Cymerman, A. (1988) Operation Everest II: Lack of an effect of extreme altitude on visual contrast sensitivity. Aviation, Space and Environmental Medicine, 59, 140-144.
- Luria, S., & Morris, N. (1988) Visual sensitivity under reduced oxygen. Groton, CT: Naval Submarine Medical Research Laboratory Report Number 1108.
- Pearson, P. G., & Neal, G. L. (1970) Operator performance as a function of drug, hypoxia, individual, and task factors. Aerospace Medicine, Feb, 154-158.
- Petjan, J. H. (1973) Neuropsychological acclimatization to high altitude. Journal of Human Evolution, 2, 101-115.
- Shukitt, B.L., Burse, P.L., Banderet, L.E., Flight, F.P., & Cymerman, A. (1988)

Cognitive performance, mood states, and altitude symptomology in 13-21* Oxygen Environments (Technical Report No. T18-88). Natick, MA: US Army Research Institute of Environmental Medicine.

Tune, G. S. (1964) Psychological effects of hypoxia: Review of certain literature from the period 1950 to 1963. Perceptual and Motor Skills, 19, 551-562.

DISCLAIMER

Naval Medical Research and Development Command, Navy Department, Research Work Unit No.61152N MR00001.01-5103. The views expressed in this article are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.

COGNITIVE PERFORMANCE AND MOOD STATES IN 13-21% OXYGEN ENVIRONMENTS

Barbara L. Shukitt and Louis E. Banderet

US Army Research Institute of Environmental Medicine
Natick, Massachusetts 01760-5007

To reduce the risk of damage from fires and loss of life in confined shipboard spaces, naval engineers have suggested that the oxygen concentration be reduced below the normal ambient level of 21% to oxygen concentrations ranging from 19 to 13% (Knight & Callahan, 1987). Air with 17% oxygen has an oxygen partial pressure (PO₂) of 129 torr which is equivalent to that at 5,600 feet, the approximate altitude of Denver, CO. Medically normal residents of that city live, work, and play without noteworthy detrimental effects. On the other hand, air with 13% oxygen has a PO₂ of 99 torr, equivalent to that at 12,600 feet, and this level of altitude has been shown to have detrimental effects on performance and mood states (Banderet, 1977; Cudaback, 1984; Fowler, Paul, Porlier, Elcombe, & Taylor, 1985; Green & Morgan, 1985; Shukitt & Banderet, 1988; Tunc, 1964).

There is an extensive literature documenting the effects of altitude on cognitive performance. The threshold altitudes for performance decrements on mental tasks and the magnitude of the impairments vary with the difficulty and complexity of the task (Cudaback, 1984). Most cognitive, motor, and affective changes occur at altitudes greater than 10,000 feet; above this altitude, motor movements are slowed, memory is less reliable, and thinking becomes confused and difficult (Fowler, et al., 1985; Tunc, 1964). Cognitive decrements also follow a specific time course at altitude. Cognitive performance can be degraded as soon as 1-6 hours after exposure, hours before the onset of altitude sickness (Banderet & Burse, 1984; Cudaback, 1984). Initial impairments are usually followed by progressive return to baseline.

Less is known about the effects of altitude on mood states. Observed behaviors and personal anecdotes suggest that ascents to altitudes above 8,000 feet cause changes in mood. The initial euphoria at higher altitudes is followed by depression and then, with time, individuals may also become quarrelsome, irritable, and apathetic. Mood changes are related to both level of altitude and duration of exposure. In one systematic study, Banderet (1977) reported mood changes, as measured by the Clyde Mood Scale, after 14 hours at 14,110 feet. A more detailed analysis of these changes (Banderet & Burse, 1988) showed that mood scores were adversely affected as early as 1-4 hours after ascent. The time course of moods is similar to the time course for symptoms of AMS. Clear Thinking, Sleepiness, Bizziness, and Friendliness were the moods adversely affected in both these studies.

The purpose of this study was to test the hypothesis that individuals can live and work in a 17% oxygen atmosphere (PO₂ = 129 torr) without experiencing appreciable decrements in cognitive performance and mood states. On the other hand, it is thought that the 13% oxygen atmosphere (PO₂ = 99 torr) will adversely affect performance and mood. A second purpose was to evaluate cognitive performance and mood for 7.5 hours in a 17% oxygen environment with a reduced pressure (PO₂ = 99 torr) for comparison with the same PO₂ in normobaric 13% oxygen.

METHODS

Subjects

Fifteen fully trained, male volunteers from the U.S. Army and Marine Corps completed this study. They ranged in age from 20 to 35 years with a mean

Assessment Instruments

ed as part of the Navy's Performance Evaluation Tests for Environmental Research (PETER) program (Bittner, Carter, Kennedy, Harbeson, & Krause, 1984; Carter & Sbisá, 1982). The computerized versions of the Pattern Comparison and Pattern Recognition tasks were used for the first time in this study; all the other tasks have been shown previously to stabilize with practice and to be sensitive to a variety of environmental stressors (Banderet, Shukitt, Crohn, Burse, Roberts, & Cymerman, 1986; Jobe & Banderet, 1986). All paper and pencil tasks were available in 15 alternate forms. The computerized tasks were developed to closely parallel their paper and pencil counterparts. A more detailed description of the cognitive tasks can be found in Shukitt, Burse, Banderet, Knight, & Cymerman (1988).

The Clyde Mood Scale (Clyde, 1963) was administered on portable computers and was used to measure subjects' moods. This scale consists of 48 adjectives, e.g., "kind", "alert", "lonely", self-rated on a four-point scale ("not at all", "a little", "quite a bit", and "extremely"). The 48 adjectives cluster into six principal mood factors - Friendliness, Aggressiveness, Clear Thinking, Sleepiness, Unhappiness, and Dizziness (Clyde, 1963).


Procedures

This experiment was part of a larger study; the research design and other aspects of the study are described in Knight and Callahan (1987) and Shukitt-Hume, Banderet, Knight, & Cymerman (1988). Extensive training and practice were given on all cognitive tasks for six to eight training days. All tasks were timed and the durations ranged from three to six minutes. Performance feedback was provided to the subjects after each practice session and extensive practice was given on all tasks to ensure that performance was stable prior to the experimental phase of the study. Subjects were also familiarized with the use of the Clyde Mood Scale during training.

Subjects were continuously confined for the fifteen day experimental

FIGURE 1

SAMPLE ITEMS FOR EACH COGNITIVE TASK

<div>ADDITION</div> <div>71 20 27 53 20 19 51 83 33 35 76 40 47 67 11</div>	<div>CODING</div> <div>NUMBER: 1 2 3 4 5 6 7 8 SYMBOL: O = U = L X = / L O , * , L / O - () () () () () () () ()</div>
<div>COMPUTER INTERACTION</div> <div>73374 MINUS 30776.9 = 58.65 PERCENT OF 41930.9 = 7398.99 DIVIDED BY 54.88 = 8897 PLUS 69194765 = 4590.84 MULTIPLIED BY 271.1 =</div>	<div>NUMBER COMPARISON</div> <div>845793858 _ 845793858 50237 _ 20237 976 _ 976 0623385 _ 0623325 239055610 _ 233055610</div>
<div>MAP COMPASS</div> <div>PVT LEE TRAVELS ON AN AZIMUTH OF 47 WHAT DIRECTION IS HE TRAVELING? _SE _NW _S _N A COMPANY IS AT GRID COORDINATE 112801. IF THEY MOVE SOUTH 1800 M WHAT IS THEIR NEW LOCATION? _064801 _130801 _112818 _112768</div>	<div>PATTERN COMPARISON</div> <div>* * * * * * * * * _ _ _ * * *</div>
<div>PATTERN RECOGNITION</div> <div></div>	

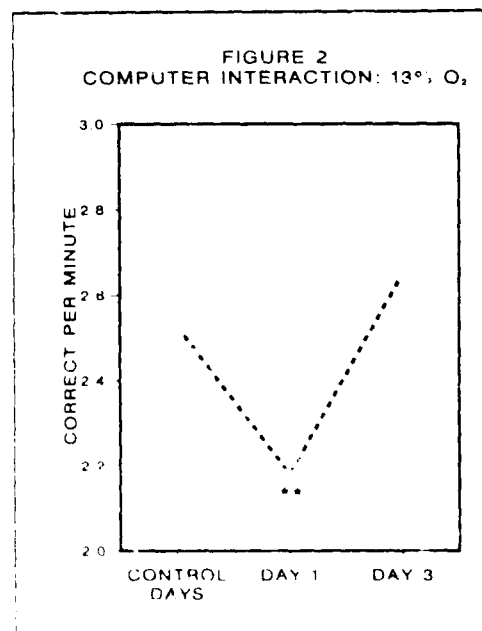
testing period to the USARIEM hypobaric chamber, where they lived for three days at each of the following oxygen concentrations: 21, 17, 21, 13, and 21%. The oxygen concentration was changed slowly every third day (e.g., from 21% to 17% to 21%) from 0100 to 0500 hours while the subjects were sleeping. All atmospheres contained $0.9 \pm 0.2\%$ carbon dioxide; the balance was nitrogen. The barometric pressure ranged from 744 to 771 torr, except for 7.5 hours on day 3 at 17% oxygen when the pressure was reduced to 576 torr (beginning at 1300 hours) to produce a PO₂ of 99 torr.

Different cognitive task batteries were administered morning and afternoon during the 30-minute sessions. During the morning session, the Computer Interaction, Pattern Comparison (paper), Pattern Comparison (computer), and the Pattern Recognition tasks were given. During the afternoon session, the Addition, Coding, Map/Compass, and Number Comparison tasks were given. The Clyde Mood Scale was also administered each afternoon after the cognitive tasks. The afternoon tasks were administered every day, but due to other project requirements, the morning tasks could only be administered on the first and third day at each oxygen concentration.

A measure of cognitive performance was derived to reflect the combined effects of changes in both rate and accuracy. Performance and mood scores were then submitted to two statistical analyses. The first was a one-way repeated-measures analysis of variance (ANOVA) to determine whether there were overall differences for the 17 or 13% oxygen condition compared to control (21% oxygen). Then, post hoc comparisons were performed with the Newman-Keuls test to identify which daily values at 17 or 13% oxygen were different than the control (21% oxygen). A significance level of $p \leq .05$ was chosen for all statistical tests. (Note: A single asterisk on the figures indicates a significant difference of $p \leq .05$ and a double asterisk indicates a significant difference of $p \leq .01$.) Greater description of the analyses are described elsewhere (Shukitt, Burse, Banderet, Knight, & Cymerman, 1988).

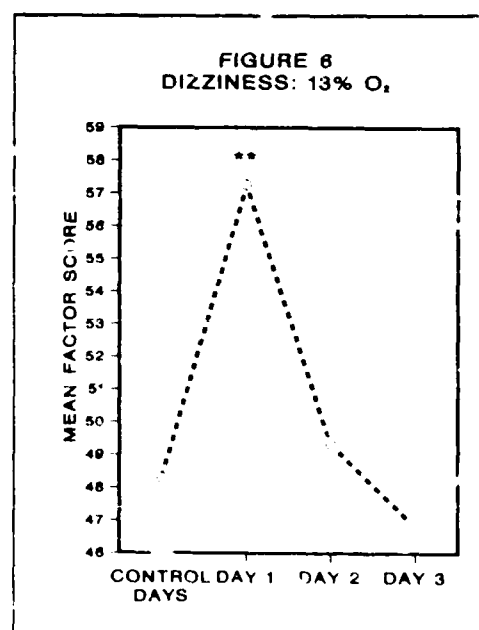
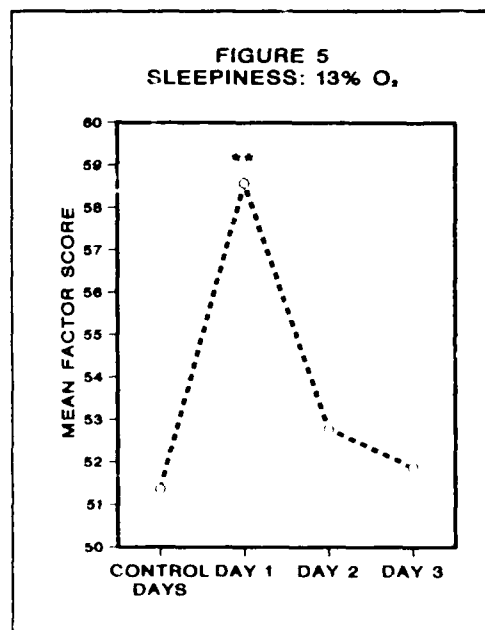
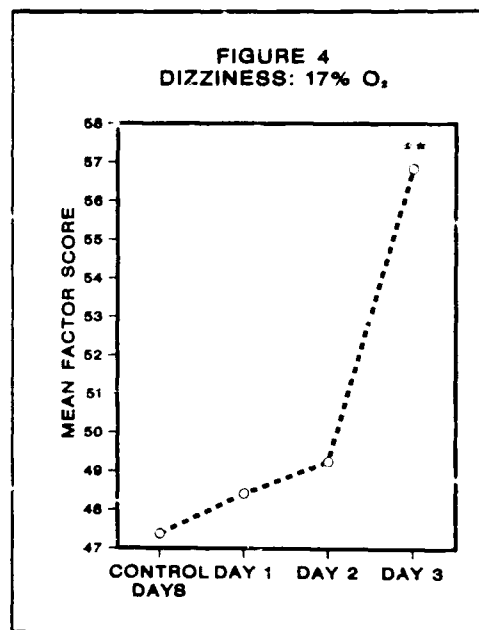
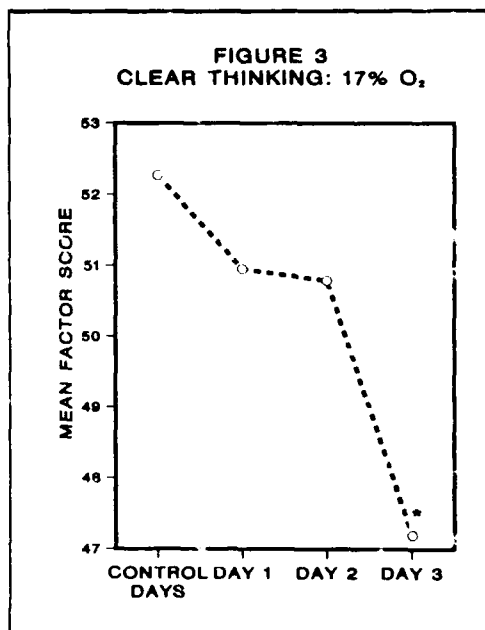
RESULTS

There were a number of significant overall effects for the cognitive tasks; Coding, Number Comparison, Pattern Comparison (paper), and Pattern Comparison (computer) at 17% and Addition, Coding, Map Compass, Number Comparison, and Pattern Comparison (computer) at 13%. No decreases in performance were observed at 17% oxygen, including day 3 when the reduction in ambient pressure occurred. Cognitive performance was decreased by 13% oxygen on only one task; scores on the Computer Interaction task were significantly less than the control value on the first day at 13% (see Figure 2). All other comparisons revealed that cognitive performance was significantly improved over the corresponding control values. These tests included Addition on days 2 and 3 at 13%, Coding on day 2 at 13%, Number Comparison on day 3 at both 17% and 13%, Pattern Comparison (paper) on day



3 at 17%, and Pattern Comparison (computer) on day 3 at 17%.

Mood scores did not change under the normobaric 17% oxygen condition. Two mood factors were affected by the reduction in pressure on day 3 at 17% oxygen; Clear Thinking was decreased (Figure 3) and Dizziness was increased (Figure 4) from control values on this day. These adverse effects resolved after the oxygen concentration was returned to 21%. Both Sleepiness (Figure 5) and Dizziness (Figure 6) scores increased on day 1 at the 13% oxygen concentration. However, by day 2, they returned to control values.



DISCUSSION

These results indicate that performance was decremented on only one task, Computer Interaction, and only on the first day at the 13% oxygen concentration. In contrast, several cognitive tasks showed gradual,

progressively enhanced performance with increased days at a specific oxygen concentration, four tasks at 17% and five tasks at 13%. The small but consistent gains in performance have been observed previously by other investigators even after extensive practice and stabilized performance (Banderet, Shukitt, Crohn, Burse, Roberts, & Cymerman, 1986; Bittner, Carter, Kennedy, Harbeson, & Krause, 1984). The fact that such gains were observed during the 17 and 13% oxygen conditions strongly suggest these conditions had little if any impact on performance.

Other studies using these tasks demonstrated performance decrements at altitudes of 13,800 to 15,500 feet (Banderet & Burse, 1984; Banderet, Shukitt, Crohn, Burse, Roberts, & Cymerman, 1986; Jobe & Banderet, 1986). Such sensitivity to altitude shown by these tasks suggests that it is likely we would have been able to measure possible performance decrements at 13% oxygen had they occurred. It is also of interest that the Computer Interaction impairment was measured at 0800 hours, three hours after the oxygen concentration in the chamber reached 13%. This result is consistent with the finding that cognitive performance may be affected by altitude as soon as 1-6 hours after ascent (Banderet & Burse, 1984; Banderet, Shukitt, Crohn, Burse, Roberts, & Cymerman, 1986; Cudaback, 1984).

We speculated why performance on the Computer Interaction task was the only one decremented by 13% oxygen. The subjects reported themselves as being more sleepy on this first day at 13% oxygen; the Computer Interaction task was the first to be given in the morning session. It may have taken the subjects longer to "get going" that morning, thus adversely affecting Computer Interaction scores without affecting tasks performed later that same day.

With respect to mood states, the 17% oxygen concentration had no adverse effect on mood states at normal barometric pressure. However, both Dizziness and Clear Thinking were adversely affected one hour after the reduction in ambient pressure on the third day at 17% oxygen. When measured after eight hours at 13% oxygen (at the same PO₂ as on the third day at 17%), Dizziness and Sleepiness scores were increased. Mood states were probably affected earlier at 13% oxygen, but this was the first administration at this condition, limiting our ability to determine at what point in time these changes occurred. The mood states of Friendliness, Clear Thinking, Dizziness, Sleepiness, and Unhappiness have been shown to be altered as early as from one to four hours after ascent to 14,110 feet (Shukitt & Banderet, 1988). Our finding of significant adverse changes after one and eight hours at 12,600 feet supports the view that the initial alteration in mood occurs early in the exposure, rather than late.

In summary, the most severe atmospheric condition tested, 13% oxygen, may produce short-term decrements in cognitive functioning and mood, lasting for less than one day. An atmosphere of 17% oxygen with a reduction in pressure could initially have an adverse effect on mood states. An atmosphere of 17% oxygen alone does not appear to produce any differences in cognitive performance or mood states, and may, therefore, be a reasonable compromise for decreasing the fire risk in confined shipboard spaces.

REFERENCES

- Banderet, L.E. (1977). Self-rated moods of humans at 4300 m pretreated with placebo or acetazolamide plus staging. Aviation, Space, and Environmental Medicine, 48(1), 19-22.
- Banderet, L.E., Burse, R.L. (1984, August). Cognitive performance at 12,600 meters simulated altitude. Paper presented at the Annual Meeting of the American Psychological Association, Toronto, Canada.

- Banderet, L.E., Shukitt, B.L., Crohn, E.A., Burse, R.L., Roberts, D.E., Cymerman, A. (1986). Effects of various environmental stressors on cognitive performance. Proceedings of the 28th Annual Conference of the Military Testing Association (pp. 592-597). Mystic, CT: US Coast Guard Academy, Department of Economics and Management.
- Bittner, A.C. Jr., Carter, R.C., Kennedy, R.S., Harbeson, M.M., Krause, M. (1984). Performance evaluation tests for environmental research: Evaluation of 112 measures (Report No. NBDL84R006 or NTIS AD152317). New Orleans, LA: Naval Biodynamics Laboratory.
- Carter, R.C., Sbisa, H. (1982). Human performance tests for repeated measurements: Alternate forms of eight tests by computer (Report No. NBDL8213003). New Orleans, LA: Naval Biodynamics Laboratory.
- Clyde, D.J. (1963). Manual for the Clyde Mood Scale. Coral Gables, FL: Biometric Laboratory, University of Miami.
- Cudaback, D.D. (1984). Four-km altitude effects on performance and health. Publications of the Astronomical Society of the Pacific, 96, 463-477.
- Fowler, B., Paul, M., Porlier, G., Elcombe, D.D., Taylor, M. (1985). A re-evaluation of the minimum altitude at which hypoxic performance decrements can be detected. Ergonomics, 28, 781-791.
- Green, R.G., Morgan, D.R. (1985). The effects of mild hypoxia on a logical reasoning task. Aviation, Space, and Environment Medicine, 56, 1004-1008.
- Jobe, J.B., Banderet, L.E. (1986). Cognitive testing in military performance research. Proceedings of the Workshop on Cognitive Testing Methodology (pp. 181-193). Washington, DC: National Academy Press.
- Knight, D.R., Callahan, A.B. (1987). The review of a research proposal to study the effects of 130 torr oxygen on submarines (NSMRL Memo Report No. 87-1). Groton, CT: The Naval Submarine Medical Research Laboratory.
- Shukitt, B.L., Banderet, L.E. (1988). Mood states at 1600 and 4300 meters terrestrial altitude. Aviation, Space, and Environmental Medicine, 59, 530-532.
- Shukitt, B.L., Burse, R.L., Banderet, L.E., Knight, D.R., Cymerman, A. (1988). Cognitive performance, mood states, and altitude symptomatology in 13-21% oxygen environments (Report No. T18-88). Natick, MA: U.S. Army Research Institute of Environmental Medicine.
- Tune, G.S. (1964). Psychological effects of hypoxia: review of certain literature from the period 1950 to 1963. Perceptual and Motor Skills, 19, 551-562.

ALTITUDE ILLNESS SYMPTOMATOLOGY IN 13, 17, AND 21% OXYGEN ENVIRONMENTS

Richard L. Burse, Charles S. Fulco and Allen Cymerman

US Army Research Institute of Environmental Medicine¹
Natick, Massachusetts 01760-5007

Douglas R. Knight

Naval Submarine Medical Research Laboratory
Naval Submarine Base New London, Groton, CT 06349-5900

Reducing the concentration of oxygen (O_2) in the air within confined spaces below its normal ambient value of 20.93% (hereafter abbreviated 21%) will decrease the flammability of combustible materials, thereby increasing fire safety. Drastically reducing the amount of O_2 in the breathing air has obvious consequences for health, mentation and physical performance; the problem is to find a level effective in reducing combustion without seriously affecting the occupants. A multitude of aerospace, submarine, and mountaineering reports suggest that effects on healthy people are rare below an altitude equivalent of 6,000 ft (16.8% O_2), but very likely in some degree above 12,000 ft (13.3% O_2). To determine the effects of prolonged exposure to reduced O_2 concentrations in the breathing air, the US Navy sponsored a joint study with the US Army into visual, cognitive, psychomotor, and affective consequences of 3-d exposures to 17 and 13% O_2 concentrations at normal sea-level ambient pressures, which included determining the incidence and severity of any symptoms of acute mountain sickness (AMS).

The symptoms of AMS, principally headache, anorexia, nausea, vomiting, insomnia, lassitude, and general malaise (1,2), appear in susceptible individuals within from 4 to 8 hours if the partial pressure of oxygen (P_{O_2}) is reduced to less than 110 torr (10,000 ft equivalent, 5). The lower the P_{O_2} below this threshold and the less time spent at intermediate P_{O_2} levels, the greater the incidence and severity of AMS symptoms. The P_{O_2} of normal ambient air (21 % O_2) at sea level pressure is 159 torr (5). The P_{O_2} of 17% O_2 is 129 torr, equivalent to ambient air at 5,600 ft (5), which will not induce AMS symptoms in normal individuals. The P_{O_2} of 13% O_2 air at sea level is 99 torr, equivalent to 12,600 ft (5), which induces mild to moderate AMS symptoms in 25-40 % of unacclimated individuals. In uncomplicated cases symptoms disappear within a few days to a week, as the body acclimates to the reduced P_{O_2} (3,4).

1. The views, opinions, and/or findings contained in this report are those of the authors and should not be construed as official Department of the Army or Department of the Navy position, policy, or decision, unless so designated by other official documentation. Human subjects participated in these studies after giving their free and informed voluntary consent. Investigators adhered to AR 70-25 and USAMRDC Regulation 70-25 on Use of Volunteers in Research. Furthermore, the protocol and procedures for the study were approved by the Committee for Protection of Human Subjects, Naval Submarine Medical Research Laboratory, Groton, CT.

The purpose of the study was to test the hypothesis that individuals can live and work in 17% O₂ (P_{O2} = 129 torr) for up to 3 d without experiencing appreciable AMS symptomatology, but not in 13% O₂ (P_{O2} = 99 torr). The effects of 17% O₂ at a pressure of 576 torr (P_{O2} = 99 torr) for 7.5-8 h were also evaluated for comparison with 13% O₂, as reduced pressures occur occasionally for that time in some submarine operations.

METHODOLOGY

Thirteen fully informed male volunteers from the US Navy and Marine Corps, ranging in age from 18-36 years (mean 23.8), completed this study. After an 8-d normoxic training period, all subjects were exposed for 3-d periods to the following conditions in a hypobaric chamber:

1. CONTROL I: P_{O2} = 159 torr (21% O₂, ambient pressure).
2. EXPERIMENTAL I: P_{O2} = 129 torr (17% O₂, ambient pressure).
(1200-1930 h, day 3, pressure = 576 torr, P_{O2} = 99 torr).
3. CONTROL II: P_{O2} = 159 torr.
4. EXPERIMENTAL II: P_{O2} = 99 torr (13% O₂, ambient pressure).
5. CONTROL III: P_{O2} = 159 torr.

Ambient pressure ranged between 744-771 torr throughout the study. O₂ concentrations were reduced by dilution with U.S.P. nitrogen. Carbon dioxide concentration was maintained at 0.9 ± 0.2%. Chamber conditions were changed slowly between 0001-0500, when subjects were asleep. Subjects were not informed of the schedule of environmental conditions, but the 13% hypoxia was quite obvious.

The presence and severity of AMS symptoms were assessed by means of the Environmental Symptoms Questionnaire (ESQ), a 67-item instrument (7) administered in card format each day just before the evening meal. Individual AMS symptoms were consolidated to produce scores for the cerebral and respiratory forms of AMS (AMS-C and AMS-R, respectively) by the weighting and summing method of Sampson *et al.* (6). By Sampson's criteria, scores exceeding a threshold of 0.6 for AMS-C and 0.7 for AMS-R are considered to be within the 95% confidence limits for the presence of that form of AMS.

Resting arterial blood O₂ saturations were determined from a direct-reading, pulse oximeter (Novamatrix model 500, Wallingford, CT), mounted on the index finger of the non-preferred hand. Determinations were made in the afternoon of the second day of each three-day exposure period, and also on the third day of the 13% O₂ exposure period for 6 subjects.

RESULTS

AMS-C and -R scores are presented in Table 1 for the three control and two experimental periods, along with the associated blood O₂ saturation levels. The subject group was divided into sick and well subgroups, based on their AMS-C scores during each experimental period. The average AMS scores and saturation values were then determined for these subgroups during each experimental period and the control periods before and after. Thus we are able to see any differences in the groups before, during, and after the experimental exposures in which they were sick or well.

Table 1: Environmental conditions, AMS scores, and blood O₂ saturations for 3-day control and experimental periods.

PARAMETER	CONTROL I		EXPERIMENTAL I			CONTROL II		EXPERIMENTAL II			CONTROL III	
	DAY 1	DAY 2	DAY 1	DAY 2	DAY 3	DAY 1	DAY 2	DAY 1	DAY 2	DAY 3	DAY 1	DAY 2
O ₂ (%)	17	21	17	17	17	21	21	13	13	13	21	21
P _B (torr)	760	760	760	760	576	760	760	760	760	760	760	760
P _{O₂} (torr)	129	159	129	129	99	159	159	99	99	99	159	159
Number: sick	0	0	0	0	2	0	0	5	2	2	0	0
well	13	13	13	13	11	13	13	8	11	11	13	13
AMS-C: sick	(0.2±0.2)	(0.2±0.2)	(0.2±0.1)	(0.2±0.2)	1.0±0.1	(0.0±0.1)	(0.0±0.1)	1.4±0.6	0.8±0.0	0.6±0.6	(0.1±0.1)	(0.1±0.1)
well	(0.0±0.1)	(0.1±0.1)	(0.1±0.2)	(0.1±0.1)	0.2±0.2	(0.0±0.0)	(0.0±0.0)	0.3±0.2	0.2±0.3	0.0±0.1	(0.0±0.1)	(0.0±0.1)
AMS-R: sick	(0.2±0.2)	(0.2±0.2)	(0.2±0.2)	(0.2±0.3)	0.7±0.4	(0.1±0.1)	(0.1±0.1)	1.0±0.6	1.2±0.6	0.8±0.2	(0.1±0.1)	(0.1±0.1)
well	(0.1±0.2)	(0.1±0.1)	(0.1±0.2)	(0.1±0.1)	0.2±0.2	(0.1±0.1)	(0.1±0.1)	0.3±0.3	0.2±0.1	0.2±0.2	(0.1±0.1)	(0.1±0.1)
S _{O₂} (torr)	n.d.	(98.6±1.1)	n.d.	(97.0±0.0)	n.d.	(98.2±0.8)	(98.2±0.8)	n.d.	92.5±1.0	93.5±0.7	(98.5±1.0)	(98.5±1.0)
	n.d.	(98.2±0.7)	n.d.	(97.3±0.9)	n.d.	(98.1±1.0)	(98.1±1.0)	n.d.	91.7±2.5	91.8±1.2	(97.8±1.0)	(97.8±1.0)

Notes: n.d. = not determined on that day.
() = mean ± s.d. computed for sick or well group, based on AMS-C score during hypoxic exposure.

During the 17% O₂ experimental period (P_{O2} = 129 torr), no subjects reported AMS-C or -R scores in the sick range until the third day, when the P_{O2} was reduced to 99 torr for 7.5 h. At the end of that period, two subjects reported AMS-C scores of a mild to moderate degree of illness (group average (\pm SD) = 1.0 ± 0.1). One of these individuals also had AMS-R. For comparison, the average values in the non-sick group were 0.2 ± 0.2 for both AMS-C and AMS-R. These scarcely differed from the control values. Resting blood O₂ saturation values on the second day at 17% normobaric O₂ were just at 97% for both groups, which represented a 1% decrease from the 98% measured during the preceding and succeeding control periods ($0.05 < p < 0.1$).

On the initial day of the 13% O₂ exposure period (P_{O2} = 99 torr), just before supper, 5 of the 13 subjects reported AMS-C scores above the threshold for illness. Four of the five also reported AMS-R scores above the illness threshold. The average scores for the sick group were 1.4 ± 0.6 for AMS-C and 1.0 ± 0.6 for AMS-R. The eight non-sick individuals, in contrast, averaged 0.3 ± 0.2 for AMS-C and 0.3 ± 0.3 for AMS-R. The sick group had improved markedly by the second day under 13% O₂; only two subjects remained ill. On the third day, the AMS-C scores for these two were less than the threshold for illness, but the AMS-R scores, although improved, still remained above threshold. The O₂ saturation values on the second day at 13% O₂ averaged $92.5 \pm 0.1\%$ for the sick group versus $91.7 \pm 2.5\%$ for the well group. Saturation measurements on the third day at 13% (6 subjects) averaged 93.5 ± 0.7 and $91.8 \pm 1.2\%$ for the sick and well groups, respectively. These results indicated no disadvantage to the sick group in their degree of O₂ saturation and also showed that both groups experienced increasing O₂ saturation with acclimation to the hypoxic environment, as expected (4). Upon return to 21% O₂, the saturation values of both groups returned to control values.

DISCUSSION

All AMS scores during the three-day control periods were well below the thresholds for illness, even those reported in subsequent control periods by formerly ill individuals. During the two days of exposure to 17% O₂ at normal atmospheric pressure (P_{O2} = 129 torr), the AMS scores remained well below illness thresholds for all subjects. After 7.5 h in 17% at reduced pressure (P_{O2} = 99 torr), two of the 13 subjects reported AMS-C; one of these also reported AMS-R. After 17 h in 13% O₂ (P_{O2} = 99 torr), five subjects had AMS-C. Of these, four also had AMS-R. After another 24 h, three of the initial five no longer suffered from AMS and the two still sick had lower scores. These two showed continued improvement on day 3, but still had scores indicative of AMS-R. This is quite similar to the experience of mountain climbers in whom a P_{O2} of 129 torr does not induce AMS in normal individuals. A P_{O2} of 99 torr, equivalent to 12,600 ft, is low enough to induce AMS in some climbers after 7-8 h and can be expected to eventually afflict about 25 - 40% of individuals staying at that altitude (2,4), most of whom will recover rapidly. Interestingly, of our two individuals with AMS for three days, it was the less severely ill who had also suffered AMS-C during the preceding hypobaric 17% O₂ exposure, not the individual more severely ill. This is not unusual; onset of illness appears to be unrelated to severity (4).

The blood O₂ saturation values obtained at rest during the 17% and 13% O₂ conditions indicated no relationship to AMS. It is worth remarking, however,

that the two most severely ill individuals at 13% O₂ had average saturation values which were numerically higher than the average of their well counterparts. Also, there was no suggestion that those later to be ill during the hypoxic exposure were so predisposed by having lower saturation values during the prior control period. An association between desaturation at rest and AMS has been shown in more severely hypoxic environments, however (2,3), probably due in part to the profound desaturation that occurs as a result of breathing disturbances during sleep (4). The current saturation measurements, taken during seated rest, cannot give any indication of the degree of desaturation experienced during sleep, of course. These quite possibly might have differed between the ill and the well groups. However, the AMS experienced after 7.5 h exposure to the hypobaric 17% O₂ environment was not influenced either by sleep or any difference in the preceding O₂ saturation. It is clear, therefore, that susceptibility to AMS must be influenced by at least one factor other than arterial desaturation.

It thus appears that about a third of the individuals exposed to 13% O₂ will experience mild to moderate AMS, but most will recover quickly. Few, if any, will suffer AMS in 17% O₂ at normal pressures, but about one-fifth will experience AMS if the ambient pressure is reduced to 576 torr for 7-8 h. Because of the similarity of AMS experience in this study to that of actual altitude exposure, it appears possible to estimate the AMS incidence at other combinations of ambient pressure and O₂ concentration from that of mountain climbers at the same P_{O₂}. This suggests that an O₂ concentration in the 14.5-15% range (P_{O₂} = 110-115 torr), equivalent to 10,000 ft, might prove acceptable from the AMS standpoint for increasing fire safety, provided that the air pressure is not reduced below that of sea-level.

REFERENCES

1. Carson, R.P., Evans, W.O., Shields, J.L. (1969). Symptomatology, pathophysiology and treatment of acute mountain sickness. Federation Proceedings, 27:1085-1091.
2. Hackett, P.H. (1978). Mountain sickness: prevention, recognition, and treatment. Albany, CA: Mountain Travel.
3. Hackett, P.H., Rennie, D., Levine, H.D. (1976). The incidence, importance, and prophylaxis of acute mountain sickness. Lancet, 11:1149-1155.
4. Hultgren, H.N. (1979). High altitude medical problems (Medical Progress). Western Journal of Medicine, 131:8-23.
5. International Civil Aviation Organization (1979). Manuel de l'atmosphère type OACI. Montreal: ICAO.
6. Sampson, J.B., Cymerman, A., Burse, R.L., Maher, J.T., Rock P.B. (1983). Procedures for the measurement of acute mountain sickness. Aviation, Space, and Environmental Medicine, 54:1063-1073.
7. Sampson, J.B., Kobrick, J.L. (1980). The environmental symptoms questionnaire: revisions and new field data. Aviation, Space, and Environmental Medicine, 51:872-877.

How Small Unit Cohesion Affects Performance

Guy L. Siebold

U.S. Army Research Institute for the
Behavioral and Social Sciences

This is a simple essay whose purpose is to articulate the nature of cohesion, of performance, and of how they intertwine. That small unit cohesion does relate to performance has been suggested by, among others, Blades, 1986; Manning and Ingraham, 1983; Oliver, 1987; Siebold, 1987a, 1987b; and Siebold and Kelly, 1988a, 1988b. But how cohesion relates to performance and what the causal mechanisms are have never been well laid out. Thus attempts to demonstrate the relationship with empirical "hard" data have met with uneven success. The problem of obtaining empirical support has been exacerbated by inadequate conceptualization of cohesion (e.g., as interpersonal attraction or confidence) and of performance as well as the difficulties of measurement.

Cohesion

To deal with cohesion, one must think in terms of the pattern of relationships between the members of a unit, between the members and the organizational segments of the unit, and between the members and the unit as a whole. These patterns or networks of relationships describe the cohesion structure of a unit. Where the relationships are strong, the structure will hold together under stress; where they are weak, the structure may disintegrate under stress. Cohesion, more properly, unit cohesiveness, may be defined as the degree to which mechanisms of social control operant in a unit maintain a structured pattern of social relationships between unit members, individually and collectively, necessary to achieve the unit's purpose (Siebold, 1987b). To deal with the relationship structure in a platoon, for example, one must examine the relationships among the thirty or so individuals and their relations with the various teams, squads, and the platoon as a whole.

It is advantageous for a soldier to have a positive relationship with another member of the platoon. In exchange for the normative and legal constraints and the other costs of the relationship such as time and resources, the soldier can receive help, protection, companionship, and other benefits. However, each additional relationship has a declining marginal utility to the soldier. Thus at some point the cost of an additional relationship begins to outweigh the direct benefits. One cannot be close friends with everybody. Additional relationships tend to become shallow and ritualistic. But where there is stability of platoon membership and the organization pays the costs of bringing the members together under common norms and goals, it is relatively cheap for the members to develop positive working

relationships with the rest of the members of the platoon. In short, it is worth investing in one another.

Over time some relationships sour, are re-established, or are built up with new members. At any given time, except perhaps during an initial "honeymoon" period, only a portion of the relationships are positive. Some are neutral; some are negative. Some members in a platoon like their squad but not another squad, or vice versa. Some members may like their platoon leader but not their platoon sergeant, or vice versa. A given pair, team, squad, or the platoon as a whole may be tightly bonded or not so tightly bonded. Nevertheless, the pattern of the relationship structure in a platoon, and the norms and other mechanisms of social order which control the various relationships, define the extent of cohesiveness in the platoon.

Performance

Small unit performance can be considered a function of individual member performance and collective performance, which are separate but feed into one another, especially in terms of motivation and efficiency. For example, if a platoon leader knows his soldiers are particularly good marksmen, he may choose to maneuver them in training to take advantage of their skill, which in turn may motivate the soldiers to gain even higher levels of proficiency. If successful, this platoon pattern is likely to be utilized in major field exercises or combat. Nonetheless, the ability of individual members to perform their tasks and of the platoon leader to plan and maneuver are relatively distinct from the ability of the teams, squads, and the platoon as a whole to interactively function to prepare for and execute their missions.

Let us examine performance in terms of human action (Goldman, 1970). If a soldier wants to do something, he carries out a series of steps or behaviors that lead to accomplishing the action. Some actions are simple and require few steps; others are more complex and require many steps. To illustrate with an example from sports, let us consider dribbling a basketball down a court and laying it up against the backboard to make a basket. A young child has the ability to perform the action of running but usually has difficulty bouncing a basketball. As he ages and with practice, he can learn how to dribble the ball and combine that action with the action of running to form the more complex, less elementary, action package of dribbling the basketball down the court. With adequate physical maturity and practice, he can put together more and more elemental actions (like combining syllables into words into sentences into speeches) such as dribbling, running, avoiding an opponent, and laying up the ball into a coherent action package of receiving a pass and dribbling down the court for a layup (Siebold, 1982).

Similarly, a soldier learns to build elemental actions into larger, constructed action packages appropriate to his wants or response to the situation. Further, he is most likely to use his strongest, quickest, and most efficient elemental actions to build the larger action package (Zipf, 1972). Teams and squads

of soldiers, especially where there is membership stability, likewise learn to perform joint elemental actions and construct them interactively into larger, more complex action packages, typically following previously defined norms or doctrine. And there is a learning curve here where it takes time and practice to learn how to efficiently construct the joint lines of action.

Just as people construct conversations over time, soldiers, leaders, battle staffs, and so forth use various elemental actions up through complex interactive action packages to construct individual and joint activity, such as fighting a battle, over an extended period. And their ability to do this depends in part on the structure of relationships, or cohesion, among the actors. The latter, for example, affects the ability of a unit of actors to sense, process, and disseminate information as well as to execute decisions (Malone, 1988).

Collective Goods

That the actions of combat soldiers in squads and platoons can be explained in terms of collective goods, especially mutual survival, has been well articulated by Kviz (1978). Essentially, collective goods are valuable, usually durable or intangible, objects which are shared or "consumed" by several or all group members. A typical example would be a bridge across a river. A key feature of some collective goods is that a number of people will benefit by its existence but that the benefits gained by any given individual will be less than the costs of his actions to obtain the good. Thus there is no aggregate incentive to produce the collectively valuable good.

As a military example, one can consider the collective good of eliminating an enemy machine gun in a pill box. All the men in a platoon would benefit by its destruction; yet the risks or costs involved in attacking it are very high for any given soldier. The ideal from an individual perspective would be for the individual to do nothing while some other members of the group overcome the enemy. This individualistic position is known as the free rider problem, in which one obtains the benefit but contributes nothing to get it. The issue then is how to achieve the collective good when it is to no one's individual profit to obtain it, i.e., eliminating the machine gun.

The solution rests with the structure of cohesiveness in the platoon. The logic has been well laid out by Coleman (1988a; 1988b) and derives from the fact that where the interests of the individuals are satisfied by the same outcome, then each has an incentive to reward the others for actions to achieve it. Thus individuals can obtain benefits from obtaining the collective good and from the other individuals. The benefits together can greatly increase the ratio of benefits to costs or may provide a net profit. The key point is that the achievement of collective goods may depend upon the ability of others to provide added benefits to the individuals producing it, and this ability to add to the rewards is dependent upon an adequate relationship structure among the actors.

The added benefits may be status, prestige, or encouragement

or reciprocal promises extending in time (i.e., informal social contracts such as "you attack the machine gun this time, and I will attack it the next time"). With stability of membership in the group, not only can there be informal reciprocal promises which extend over time, but there can be value to high status in the group and a division of labor where members tend to volunteer for those missions or actions at which they are better and are motivated to get better at them. In short, it is the closed network of positive relationships (cohesion) which facilitates the attainment of collective goods which benefit or are mandatory for the aggregate of individuals but for which otherwise it is in no individual's personal net interest to contribute towards.

Symbolism

The continual operation of reciprocal exchanges of the same content by a stable network of actors is not efficient so that the practices tend to become subject to a "norm" sanctioned by the network. The set of norms developed, either prescriptive or proscriptive, transfers control over actions from the actor to those affected by it (Coleman, 1988b). This external constraint is experienced as emanating from the reified network; the soldier does not want to let his buddies down. The reified network, symbolically experienced as buddies, the team, the squad, and/or the platoon, then becomes both a source of constraint over the individual's actions as well as a source of encouragement and surplus value for the collective good such that it can elevate the soldier above his ordinary powers (Jones, 1986). The strength of the symbolic entity to constrain or elevate the individual is dependent on its trustworthiness and density of mutual obligations. The latter is a function of the need for help from the network, alternate sources of help, the cultural significance of seeking help, logistics of help, and the degree of closure among the network of actors (Coleman, 1988b). For example, the dense, trusting relationships among soldiers in many American small combat units in Vietnam resulted in some soldiers electing for a second tour with their units rather than trade their "family" for the less dense, fragmented networks they anticipated back in the "real world."

It is commonly observed that soldiers fight more for their buddies (networks) than for some abstract political ends. The importance of interpersonal bonds to soldiers is indicated by the importance soldiers attach to their rituals (e.g., handshakes or salutes) and identification with symbolic groups (e.g., the squad or the platoon or higher), if there is a strong, closed network. Further, the symbolic groups, affirmed through public rituals denoting fellowship and communion, represent the institutionalized relationships of the collective good and the power of the social capital they embody. The symbolic entities are capable of both inspiring confidence and demanding subjugation to the collective good. Thus it is not surprising to find soldiers adopting a totemic symbol to stand for their particular group (network) to imbue it with spiritual meaning beyond any official designation.

The importance of symbolism in terms of performance is that

it permits the control of soldiers through internalized norms and meanings and can elevate soldiers to actions above and beyond the call of duty or their personal interests. The catch is that it is dependent upon the network of positive relationships, the cohesion, in the echelon unit. The dud, the slug, the soldier who does not have positive relationships with his fellows soldiers is a loss to his group; but the leader in the chain of command who does not have positive relationships with his men may separate his whole subelement from the larger collectivity. Thus a critical function of a leader is to provide continuity in the network by positive linking relationships with his subordinates and his superordinate leaders. Particularly when there is not a charismatic superordinate leader, the symbolic entity or network ends in an organizational pyramid with the first non-positive relationship up the chain. The lower the level at which the first non-positive relationship exists, the less social capital that is available and the lower the level of symbolic entity with which soldiers will identify. Hence, group performance will tend to be strong or weak depending on how high up the chain the network of positive relationships exists and tend to maximize the collective good, such as survival, at the corresponding level.

Conclusions

The above is a way of looking at cohesion and performance which provides a motive for the generation of cohesion and an explanation of how it is developed, sustained, and related to types of performance. Further it provides a way to look at group development over time in interaction with performance. It also allows us to re-examine the literature in a systematic way and suggests directions for future research. Of particular interest is the verification of the conditions under which one would expect to find a significant cohesion-performance relationship. For example, in "old" groups with low (soldier and leader) turbulence and good leadership, one would expect the group to possess a high degree of cohesiveness, be capable of performing relatively complex action packages, have an articulated prestige/status structure based on commonly known and accepted norms, and have developed a hard core of "gung ho" soldiers. The military performance of such a group should be outstanding.

References

- Blades, Jon W. 1986. Rules for Leadership. Fort Lesley J. McNair, Washington, DC: National Defense University Press.
- Coleman, James S. 1988a. "Free Riders and Zealots: The Role of Social Networks." Sociological Theory, 6, Spring: 52-57.
- Coleman, James S. 1988b. "Social Capital in the Creation of Human Capital." American Journal of Sociology, 94, S95-S120.
- Goldman, Alvin I. 1970. A Theory of Human Action. Englewood Cliffs, NJ: Prentice Hall.

- Jones, Robert Alun 1986. "Durkheim, Frazer, and Smith: The Role of Analogies and Exemplars in the Development of Durkheim's Sociology of Religion." American Journal of Sociology, 92, No. 3, 596-627.
- Kvitz, Frederick J. 1978. "Survival in Combat as a Collective Exchange Process." Journal of Political and Military Sociology, 6, Fall: 219-232.
- Malone, Dandridge M. 1988. "The Technology of Teamwork: Military Applications of Teamwork Research." Technical Report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences (29 June) (forthcoming).
- Manning, Frederick J. and Larry J. Ingraham 1983. "An Investigation into the Value of Unit Cohesion in Peacetime." Report WRAIR NP-83-5. Washington, DC: Walter Reed Army Institute of Research (November).
- Oliver, Laurel W. 1987. "The Relationship of Group Cohesion to Group Performance: A Research Integration Attempt." Paper presented at the Annual Meeting of the American Psychological Association, New York, NY (August).
- Siebold, Guy L. 1982. "Conversation Sequencing and Joint Lines of Action." Paper presented at the 77th Annual Meeting of the American Sociological Association, San Francisco, CA (6-10 September).
- Siebold, Guy L. 1987a. "Bonding in Army Combat Units." Paper presented at the Southern Sociological Society Annual Meeting, Atlanta, GA (9-12 April).
- Siebold, Guy L. 1987b. "Conceptualization and Definitions of Military Unit Cohesiveness." Paper presented at the 95th Annual Convention of the American Psychological Association, New York, NY (28 August-1 September).
- Siebold, Guy L. and Dennis R. Kelly. 1988a. "The Impact of Cohesion on Platoon Performance at the Joint Readiness Training Center (JRTC)." Technical Report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences (June) (forthcoming).
- Siebold, Guy L. and Dennis R. Kelly. 1988b. "Development of the Platoon Cohesion Index (PCI)." Technical Report. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences (July) (forthcoming).
- Zipf, George Kingsley 1972. Human Behavior and the Principle of Least Effort. New York: Hafner Publishing Company (1949).

The Relationship between Leadership Competency Ratings and Platoon Cohesion

Dennis R. Kelly

US Army Research Institute for the
Behavioral and Social Sciences

Previous research has indicated that the leadership exhibited by platoon leaders in a tactical environment is related to the cohesiveness of the platoon (e.g., Siebold & Kelly, 1987, 1988b; Twohig & Tremble, 1987). These studies have utilized external raters to evaluate a leader's competencies. Another approach which adds validity to this research, is to have the platoon membership rate their leaders on the leader competencies. This paper will present the findings of research which examined the relationship between the ratings made by platoon members on their platoon leader, platoon sergeant and squad leaders on seven leadership competencies and the reported level of cohesion in a platoon. Additionally, results will be presented which examine this relationship controlling for the effects of morale.

The theoretical framework of small unit cohesion and its measurement have been elaborated previously. Briefly, cohesion is conceptualized as the structured pattern of social relationships between unit members, individually and collectively, necessary to achieve the unit's purpose. The framework includes bonding between first term soldiers (horizontal bonding), between first term soldiers and their leaders (vertical bonding), and between all the soldiers and their unit (organizational bonding). Each level of bonding can be sub-divided into affective and instrumental components (Siebold 1987a, 1987b; and Siebold and Kelly 1988a).

Method

Subjects: Data were collected from thirty-two (32) light infantry platoons from eight companies of two battalions of the same brigade. Only the data from the junior enlisted soldiers of ranks E1-E4 were used for analyses. The total individual sample was 316 soldiers.

Instrument: A questionnaire was used, comprising 150 items, with the last 41 items being a 20 item short form (PCI) measure of cohesion, and 21 items to measure seven leadership competencies for three platoon leadership positions. The PCI has established psychometric properties (Siebold & Kelly, 1988a). The PCI is a twenty item instrument that utilizes five point scales (scored as from 0-4 in this paper, with higher numbers indicating greater cohesion). Because of the similarity of response patterns, the ten cohesion scales that are derived from the PCI were consolidated for presentation purposes in this paper into five scales. The horizontal bonding (HB) scale represents the combined affective and instrumental bonding among peers (E1-E4 pay grades) scales. The vertical bonding (VB) scale is comprised of the affective bonding among leaders (E5-02

The opinions expressed in this paper are those of the author and do not necessarily reflect the position or policy of the U.S. Army Research Institute or the Department of the Army.

pay grades) scale, as well as the affective and instrumental vertical bonding between first termers and leaders scales. The organizational bonding (OB) scale represents the combination of the organizational pride, needs, goals and rule clarity scales. The organizational bonding, values scale items were separated to form the leader (OB-A, LV) and first termers values scales (OB-A, FTV) for the analyses.

The seven competencies (and the actual examples that were provided as definitions of the competencies in the questionnaire) that soldiers rated their leaders on were: planning (quality of plans, timeliness of plans, subordinates are consulted), communicating (speaks and writes clearly and to the point; is open; listens well), motivating (instills desire to do needed tasks well and accomplish mission), tactical and technical matters (knows tactics, weapons, equipment; uses troop leading procedures), building cohesion (trusts and develops subordinates; looks out for his men; encourages teamwork; and is a good role model), supervision (subordinates are informed of expectations and standards; leader provides feedback and takes corrective action; leader allows some flexibility for subordinates--doesn't over-supervise), and decision making (makes timely decisions; uses sound judgment; adjusts to the situation; is bold and innovative when appropriate; consults others as needed). The competencies were rated on a four point scale (scored as 0-3), from excellent (=3), to good (=2), to only fair (=1), to poor (=0). A score of 1.5 is considered the midpoint of the scale and indicates a competency is between "only fair" and "good". The competencies were selected from a larger set of key leader competencies that the Center for Army Leadership has generated and synthesized. The seven chosen were considered the most pertinent to cohesion.

The item in the questionnaire, "How high is the morale in your platoon." was used as a "control" variable in the partial correlations. This item was rated on a six point scale ranging from 0 (extremely low) to 6 (extremely high).

Procedures: The questionnaire was administered to soldiers in a classroom setting, one company at a time.

Results and Discussion

The means of the seven leader competencies for each of the three platoon leadership positions are provided in Table 1. Also provided is the average rating across each competency and leadership position. With the exception of the average rating for tactics (ranging from 1.68-1.89 for the three positions), the remaining six competencies were similar to each other in size and fall between the only fair to good rating for each leadership level. Although the seven mean competency ratings are similar in size for each position across platoons, there was sufficient variation noted within a platoon to indicate that soldiers are able to discriminate a given leader's effectiveness. The SL position is rated as highest across each of the seven competencies. The PL and PS were rated similar to each other but were both rated as lower than the SL. While the SL ratings may be an accurate assessment of their competencies, one might question whether the average SL would be any more proficient at the competencies than his more experienced PS. The higher SL ratings probably have to do with the perceptions of the raters themselves and their proximity to the three leadership positions and functions. The soldiers rating the leaders on the competencies are of E1-E4 rank, and their closest supervisor/leader is the

Table 1

Mean Competency Ratings of Platoon Leadership Positions

Position	Leader Competencies							Average
	Plan	Comm	Motv	Tact	Chsn	Sprv	Dcsn	
PL	1.53	1.56	1.41	1.68	1.45	1.48	1.51	1.52
PS	1.50	1.51	1.44	1.80	1.46	1.52	1.59	1.54
SL	1.74	1.70	1.67	1.89	1.67	1.75	1.76	1.73
Average	1.59	1.59	1.51	1.79	1.52	1.58	1.62	1.60

Note. PL=platoon leader; PS=platoon sergeant; SL=squad leader.
 Plan=planning; Comm=communicating; Motv=motivating;
 Tact=tactics; Chsn=cohesion building; Sprv=supervision;
 Dcsn=decision making; Scale ranges from 0=poor, 1=only fair,
 2=good, 3=excellent. N=32 platoons.

SL. They have more of an opportunity to witness the performance of the SL than they do of either their PS or PL. As well, soldiers may be "protecting" or "rewarding" the leader that is closest to them by giving slightly inflated ratings.

Table 2 presents the simple and partial correlations (controlling for morale) between the overall competency ratings for each of the platoon leadership positions and the reported level of cohesion in the platoon. Overall leader competency ratings were calculated for each platoon by adding the seven competency ratings and dividing by seven for each of the three leadership positions.

Examining the simple correlations in the table reveals that the ratings of the leader competencies have the strongest relationship with the VB scale. This relationship exists at each of the three leader positions, with the SL position producing the largest correlation (.74). However, the leader competencies also have strong correlations with the OB and OB-A, LV scales. Again, in the OB scale the SL has the highest correlation (.77). The size of the correlations between the OB-A, LV scale and the competency ratings for the three platoon leadership levels is approximately equal, with the PS having the largest effect (.59). The HB and OB-A, FTV scales are slightly, although significantly correlated with the SL position and not significantly related to the PS or PL positions. The OB-A, LV scale is not significantly related to the competency ratings for any of the leadership positions. In terms of the leader positions themselves, in nine out of the ten possible comparisons of the SL to the PS or PL, the competency rating of the SL has the largest correlation with the cohesion scales. The correlations of the PS and PL competencies ratings with the cohesion scales are approximately equal in strength.

Table 2

Simple and Partial Correlations of Competency Ratings Received
by Platoon Leadership Positions with Cohesion

Cohesion scales	Overall leader competency ratings					
	Simple correlations			Partial correlations		
	PL	PS	SL	PL	PS	SL
HB	.27	.11	.48	.08	.00	.29
VB	.65	.60	.74	.51	.62	.51
OB	.44	.38	.77	.20	.29	.58
OB-A,FTV	.21	.19	.29	.03	.09	.05
OB-A,LV	.50	.59	.54	.29	.58	.18

Note. HB=horizontal bonding; VB=vertical bonding; OB=organizational bonding; OB-A,FTV= organizational bonding, first term values: OB-A,LV=organizational bonding, leader values. PL=platoon leader; PS=platoon sergeant; SL=squad leader. Correlations of .29 or above are significant at the .05 level or greater. Partial correlations control for the "morale" of the platoon. N=32 platoons.

When the effects of morale are controlled for, a pattern similar to the simple correlations emerges. However, there is a reduction in the size of the correlations. The average decrease in size of the correlations is .16. While this decrease is notable, it does not greatly diminish the major relationships and suggests the relationship between the leader competency ratings and cohesion is not a spurious one. Only two of the fifteen possible simple correlations were not significant as partial correlations. The strength the relationship between the competency ratings of the leader with VB scale is maintained. While there is a decrease in the size of the correlations, the OB and to a lesser extent the OB-A,LV scales maintain similar patterns of correlations with the competency ratings. The competency ratings of the SL maintain a distinct relationship with the OB of the platoon. Also, the PS competency ratings are now more uniquely associated with the OB-A,LV scale. The influence of morale is greatest on the SL and PL competency ratings across the five cohesion scales used. The PS correlations with the cohesion scales were least influenced by the morale of the platoon. The competency ratings of the PS now emerge as having the strongest correlations with the VB and OB-A,LV scales.

Table 3 presents the correlations and partial correlations (controlling for morale) between the seven leader competencies and the cohesion scales. The competencies represent the average ratings for the three leadership positions in a platoon. Similar to table 2, the simple correlations indicate that the competencies are most strongly related to the VB, OB, and OB-A,LV scales, with the VB scale correlated the highest across each of the seven competencies. In other words, the leader competency ratings are correlated highest with those aspects of cohesion which are most sensitive to the effects of the leader.

Table 3

Correlations of Competencies with Cohesion

Cohesion scales	Leader competencies													
	Simple correlations							Partial correlations						
	Plan	Comm	Motv	Tact	Chsn	Sprv	Dcsn	Plan	Comm	Motv	Tact	Chsn	Sprv	Dcsn
HB	.39	.14	.31	.39	.37	.43	.30	.16	-.12	.09	.21	.16	.29	.10
VB	.77	.68	.81	.79	.81	.78	.75	.59	.50	.72	.70	.69	.72	.65
OB	.65	.60	.58	.58	.68	.54	.61	.41	.38	.35	.37	.48	.35	.42
OB-A,FTV	.17	.13	.29	.26	.39	.42	.26	-.09	-.09	.10	.08	.22	.29	.08
OB-A,LV	.65	.57	.67	.60	.66	.68	.60	.40	.34	.49	.41	.44	.56	.41

Note. HB=horizontal bonding; VB=vertical bonding; OB=organizational bonding; OB-A,FTV= organizational bonding, first term values: OB-A,LV=organizational bonding, leader values. Plan=planning; Comm=communicating; Motv=motivating; Tact=tactics; Chsn=cohesion building; Sprv=supervision; Dcsn=decision making; PL=platoon leader; PS=platoon sergeant; SL=squad leader. Correlations of .29 or larger are significant at the .05 level or greater. Partial correlations control for the "morale" of the platoon. N=32 platoons.

The cohesion building, motivating, and supervision competencies are significantly related to all of the cohesion scales. The tactics, decision making, and and planning competencies are significantly related to four of the five scales and the communication competency is related to the three cohesion scales most related to leadership (VB, OB, OB-A,LV).

The partial correlations reveal the same patterns that were seen in Table 2 and follow the patterns of the Table 3 simple correlations, but with reduced strength. The average decrease in the size of the correlations is .18. Correlations that involve either the VB scale or the "supervision" leader competency are the least impacted by the effects of morale.

Conclusion: There is a strong relationship between the competency of the platoon leadership and the level of both vertical and organizational cohesion in the platoon. While each leadership position is important in this relationship, the SL position appears to be most strongly related to the organizational bonding in the unit followed by the PS and then the PL. Only the SL competency ratings were related (significantly) to the level of horizontal bonding (between peers) in the platoon. The PS maintains the strongest correlations with the VB in the platoon (when the effects of morale have been controlled for) followed by the SL and PL. The seven leader competencies rated are approximately of equal importance in relationship to the cohesion of the platoon with the supervision and cohesion competencies having the strongest relationship to the level of cohesion in the platoon.

References

- Siebold, Guy L. (1987a, April). Bonding in army combat units. Paper presented at the Annual Meeting of the Southern Sociological Association, Atlanta, GA.
- Siebold, Guy L. (1987b, August). Conceptualization and definitions of military unit cohesiveness. Paper presented at the Annual Convention of the American Psychological Association, New York, NY.
- Siebold, Guy L. & Kelly, Dennis R. (1987, October). The impact of unit cohesion on unit performance, morale, and ability to withstand stress: A field exercise example. (ARI Working Paper #87-13). Alexandria, VA: US Army Research Institute.
- Siebold, Guy L. & Kelly, Dennis R. (1988a, July). Development of the Platoon Cohesion Index (PCI) (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute.
- Siebold, Guy L. & Kelly, Dennis R. (1988b, June). The impact of unit cohesion on platoon performance at the Joint Readiness Training Center (JRTC) (ARI Technical Report). Alexandria, VA: US Army Research Institute.
- Twohig, Paul T. & Tremble, Trueman R. (1987). Home station determinants of the platoon leader-platoon sergeant relationship in a tactical environment: Focused rotation at the National Training Center (ARI Working Paper #87-09). Alexandria, VA: U.S. Army Research Institute.

COMBAT READINESS: PREPARING SOLDIERS
FOR THE STRESS OF BATTLE

Major W.R. Wild

Canadian Forces Personnel Applied Research Unit
Willowdale, Ontario, Canada

Introduction

The role of narcissistic defences in mediating stress was discussed by Masserman (1955) and interpreted in the context of combat by Shaw (1981). According to the narcissistic defence theory, the soldier facing battle draws strength initially from the "invincible self". As the going gets tough, the soldier draws strength from the "omnipotent leader" (leadership) and, in a final attempt to retain control, from the presence of comrades (cohesion). Teichman (1977) demonstrated this theory when he showed that in the initial stages of combat, little palliative interaction takes place between soldiers. As combat intensifies however, soldiers turn first to leaders and finally to the old soldiers among them for reassurance. Other researchers (for example, Grinker, 1945; Marshall, 1947; Stouffer et al., 1949; Kellett, 1982; Gal, 1986) have documented the importance of cohesion, morale, and leadership in maintaining combat effectiveness in the face of extreme stress.

In a 1986 study conducted by the author in Cyprus (Wild 1987), it was hypothesized that soldiers in a stressful situation (one of high environmental demand) would report no more stress than soldiers in an unstressful situation (one of low environmental demand). It was suggested by Masserman (1955) and Shaw (1981), and demonstrated by Teichman (1977), that soldiers in a high environmental demand situation readily accept the support of peers and leaders and by so doing add to their means of coping with stress. As stress occurs when environmental demand exceeds the organisms ability to cope, the addition of these coping mechanisms delays the onset of stress. Thus, in a situation of high environmental demand, stress is ameliorated by the cohesion that is created by the high environmental demand in the first place. That is, soldiers in high environmental demand situations (combat, near combat) should report stress levels lower than would be expected given the nature of their employment because cohesion with peers and leaders reduces the individual's level of stress.

Method

Three hundred and six junior soldiers (corporals/privates) in two experimental groups (Infantry Rifle Companies patrolling the Green Line in Cyprus) and three control groups (an Administrative Support Company in Cyprus, and two Infantry Rifle Companies in Canada) participated. Stress was measured using the Self-Analysis Questionnaire, STPI Form X - 1 (Spielburger, Gorsuch, & Lushene, 1970) and a survey constructed by the

The views and opinions expressed in this paper are those of the author and not necessarily those of the Department of National Defence.

author. Vertical cohesion was measured by the General Attitude to Institutional Authority Questionnaire (Rigby, 1982), an instrument that indicates the soldier's attitude to his leaders. Cohesion was measured by a questionnaire composed of items adapted from Seashore (1954) and Pepitone and Kleiner (1957). A number of job factors were also investigated.

Results

Multivariate Analysis of Variance and Tukey's "honestly significant difference" test of means revealed limited support for the hypothesis. Quite unexpectedly, the two groups that differed most proved to be the groups that were expected to be most similar; the two Cyprus line companies. As Figure 1 shows, group 4 received scores similar to those received by the control groups on the major effects, while the other experimental group, group 5, scored higher on the stress scale and lower on the attitude to authority scale (alpha less than .05). A comparison of these two groups

SAME LETTER: NO SIGNIFICANT DIFFERENCE					
	10			A	
		AB	AB	AB	*
S	8	*	*	*	
C					B
O	6				*
R					A
E	4	B	B	B	*
		*	*	*	
	2	A	A	A	A
		*	*	*	*
		COHESION			
		1	2	3	4
		GROUP			

Figure 1. Group mean scores for main effects.

(Figure 2) showed significant (alpha less than .05) differences in attitude toward authority, intensity of stressors and attitudes toward co-workers and supervisors, with group 5, the high stress company, characterized by higher reported stress, higher reported environmental demand, greater dislike of co-workers and a less favourable opinion of supervisors. Scores on cohesion and work were also lower for group 5, although the differences failed to reach statistical significance (alpha less than .05). Discussions with personnel in both companies revealed disciplinary problems in group 5 and an unusually high number of disciplinary charges being laid by leaders.

Discussion

Despite the fact that the two experimental groups were operating in the same environment and reported no difference in the nature of the work, group 5 reported higher environmental demand. As the environment external to the companies was the same (UN peacekeeping in Nicosia, Cyprus), the environment reflected by the self-report measure must have been the internal environment: the command or organizational climate. JDI scores on the co-worker and supervisor scales support this explanation as does the observation that disciplinary problems in group 5 were greater than those in group 4. JDI, co-worker, and supervisor scale scores; differences in

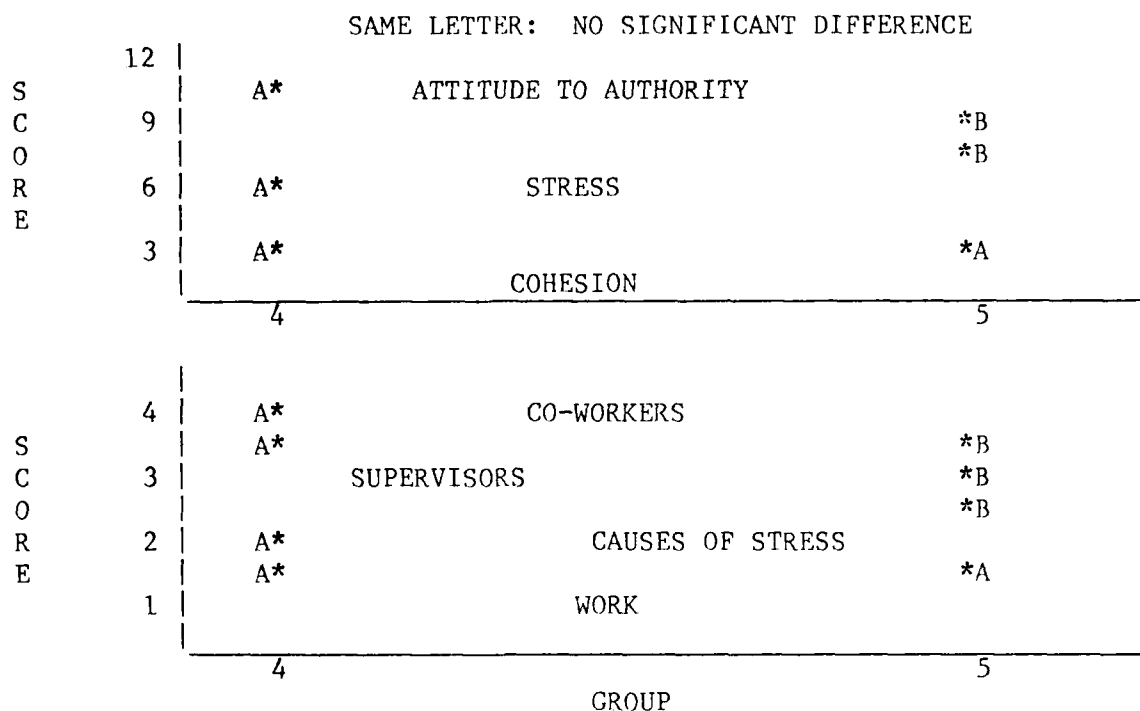


Figure 2. Comparison groups four and five.

disciplinary atmosphere; as well as the difference in scores on the cohesion index (albeit not a statistically significant difference) indicate that group 5 had less cohesion than group 4. If this is so, it suggests that vertical and horizontal cohesion are not necessarily fostered by conditions of high environmental demand. It is possible that cohesion, or the lack of it, is exacerbated by high environmental demand, with cohesive groups (e.g., group 4) pulling together when under stress and non-cohesive groups (e.g., group 5) pulling apart.

Combat Readiness

It became obvious from the Cyprus experience that the combat stress phenomenon is a complex one that does not yield readily to investigation. In order to overcome this complexity, researchers at the Canadian Forces Personnel Applied Research Unit (CFPARU) set out to develop a model to anchor further study. In the course of this development, it was realized that taken in isolation, combat stress is of little importance to the combat commander. Tactics are at the core of the combat commander's professional life. Psychological breakdown is not a consideration on the peacetime practice field and so is easily ignored. Combat stress becomes important only when it is perceived that it degrades combat effectiveness. The minimization of combat stress is not the commander's ultimate goal, the maximization of combat effectiveness is.

Unfortunately, for researchers, it is impossible to tell how combat effective a unit is unless you commit it to combat, at which time, if the unit is not effective, it is too late to do anything about it. Not being able to measure combat effectiveness, we must strive to predict combat effectiveness. Predicted effectiveness is what we term readiness: Combat

readiness, therefore, is merely predicted combat effectiveness. The model we have developed is a blueprint for all of the human dimension factors that must be considered when preparing a formed body of individuals for war. If all of the components of the model are addressed satisfactorily, the unit will be ready: from a human dimensions standpoint.

A Model of the Human Dimension of Combat Readiness

The components of the combat readiness model (Figure 3) are divided into antecedent variables (aptitude, combat proficiency, understanding of task, motivational environment and motivation), and mediating variables (leadership) (Wild, 1988). In order for a unit to be effective, the members must have the aptitude to learn the tasks they will be called upon to perform; they must, through instruction and practise, gain the knowledge, skills, and physical abilities required to make them combat proficient; and they must know what their task is, including the nature of their enemy.

The environment in which they work as individuals, small group members, and members of a larger system must foster combat readiness. As individuals, members must have the high level of personal morale that stems from knowing that if injured they will receive the best available medical attention, and that family at home are being cared for. They must have the professional morale that comes from confidence in equipment, training, and tactics; pride in self and unit; knowledge of what is expected of them; and

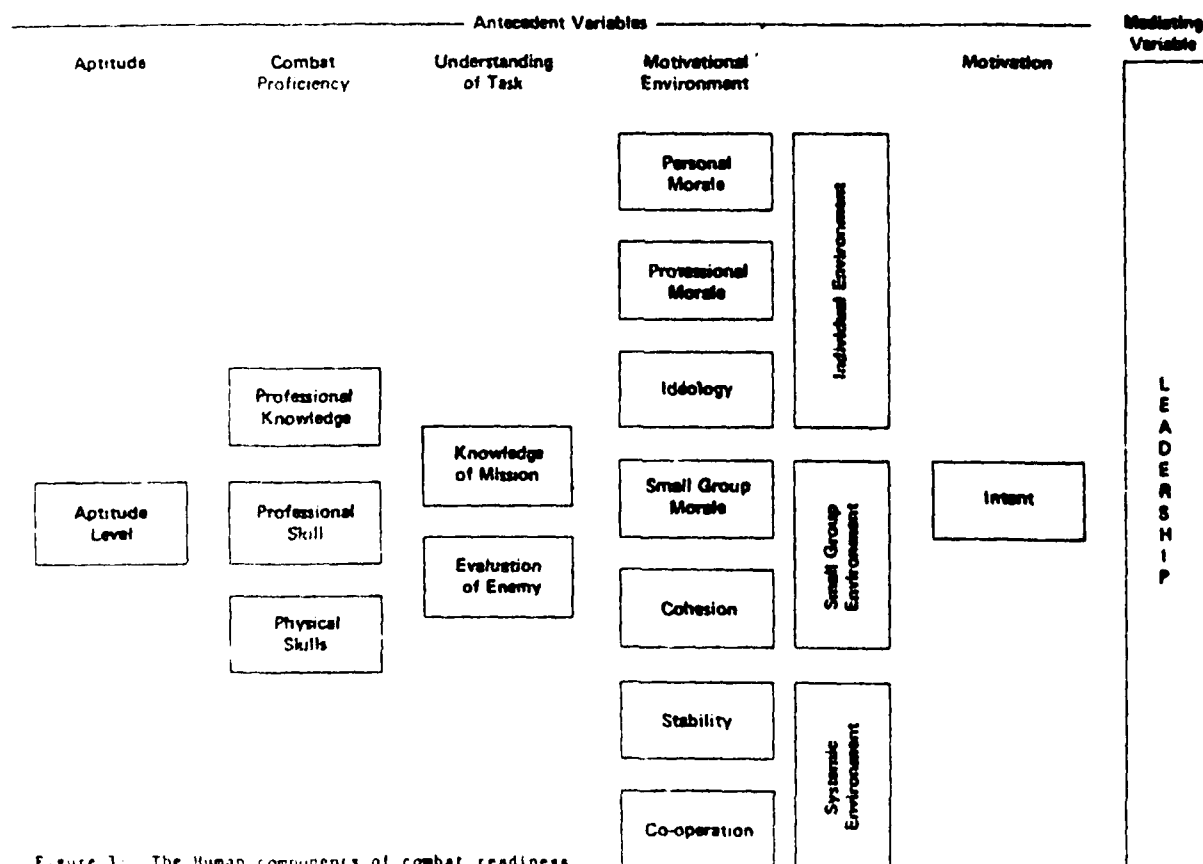


Figure 3: The Human components of combat readiness.

the belief that the organization is concerned with the individuals under its care. Individuals must also be comfortable with the ideology for which they are fighting, a component that has occasionally been absent in both western and eastern armies in the recent past.

The group also plays a crucial role in ensuring readiness. Group morale (the feeling of well-being which comes from being part of a group) and group cohesion (that which binds group members together as one) must be strong, for in the end, it is groups as small as two soldiers (or sailors, or airmen) that break down under the stress of battle. Lastly, the system must support the individual and the group. Only by remaining with one another can a body of soldiers form a cohesive group. Posting and training policies must not create the systemic instability that ensures that a group is never together long enough to become cohesive. Just as cohesiveness is important among members of a small group, cooperation among units in a formation is important. For a large unit to be combat ready, the smaller units of which it is composed must cooperate.

The last antecedent variable is intent. If all of the previous variables are present in sufficient quantity, the members of the unit should be combat ready. Unit members will not be combat ready if, in spite of everything, they have no intention of fighting.

Mediating every other variable is leadership. Leadership is the most important variable, for the leader is inextricably linked to every other component of the model. From training to the individual's final decision to fight or cower, leadership plays a part.

Although the model outlined above is labelled combat readiness, it addresses combat stress as well. It was pointed out earlier that if all of the components of the model are addressed satisfactorily, the unit will be ready for combat - from a human dimensions point of view. It will be ready because it is implicit in the model that if the components are addressed satisfactorily, combat stress will be minimized: in effect, combat readiness is the flip side of combat stress. The more ready the unit, the better able its members are to cope with stress.

Conclusion

Although the Cyprus study did not lend support to the hypothesized relationship between stress, leadership and cohesion, it did serve a useful function; it revealed combat stress to be somewhat narrow in focus, cumbersome to conceptualize for research purposes and of little interest to the combat commander. For these reasons, the focus at CFPARU was changed from combat stress to combat readiness. Although the emphasis has changed, combat stress has continued to play a central role, for each of the components of the combat readiness model is as much a variable of combat stress as it is of combat readiness. In effect, combat readiness is the "flip side" of combat stress; the same phenomenon addressed from a different angle, an angle which we think will be more readily accepted by the Combat Commander.

REFERENCES

- Gal, R. (1986). Portrait of the Israeli Soldier. New York: Greenwood Press.
- Grinker, Lt. Col. Roy R., & Spiegel, Major John P. (1945). Men Under Stress. Philadelphia: Blakiston.
- Kellett, Anthony (1980). Combat Motivation (Research Report R77). Ottawa, Canada: Operational Research and Analysis Establishment.
- Kendall, P.K., Finch, A.J., Auerbach, S.M., Hooke, J.F., & Mikulka, P.J. (1976). The State-Trait Anxiety Inventory: A systematic evaluation. Journal of Consulting and Clinical Psychology, 41, 406-412.
- Marshall, S.L.A. (1947). Men Against Fire. New York: William Morrow.
- Masserman, J. (1955). The practice of dynamic psychiatry. Philadelphia: W.B. Saunders.
- Pepitone, A., & Kleiner, R. (1957). The effects of threat and frustration on cohesiveness. Journal of Abnormal and Social Psychology, 54, 192-199.
- Rigby, K. (1982). A concise scale for the assessment of attitudes towards institutional authority. Australian Journal of Psychology, 34, 195-204.
- Seashore, S.E. (1954). Group cohesiveness in the industrial work group. Ann Arbor: University of Michigan.
- Shaw, M.E. (1981). Group dynamics. New York: McGraw-Hill.
- Speilberger, C.D., Gorsuch, R.C., & Lushene, R.E. (1970). Manual for the State-Trait Anxiety Inventory. Palo Alto, California: Consulting Psychologists Press.
- Stouffer, S.A., Lumsdaine, A.A., Lumsdaine, M.H., Williams R.M., Smith, M.B., Janis, I.L., Star, S.A., & Cottrell, L.S. (1949). The American soldier. Princeton: Princeton University Press.
- Teichman, Moir. (1977). Affiliative behaviours among soldiers during wartime. British Journal of Social and Clinical Psychology, 16, 3-7.
- Wild, W.R. (1987). Mediators of stress in an army unit. Unpublished Master's thesis, University of Manitoba, Winnipeg, Manitoba.
- Wild, W.R. (1988). Proposal for studying the human dimension of combat readiness (Technical Note 5/88). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

Measuring Leader Performance ¹ in Simulated Combat in the Field

Paul T. Twohig and Trueman R. Tremble

U. S. Army Institute for the
Behavioral and Social Sciences

The Center for Army Leadership (CAL), the U.S. Army's proponent for leadership doctrine, has had continuing interest in developments for improving leadership performance effectiveness. This interest was reinforced by the 1987 Leader Development Study which identified the need to establish the leadership performances and standards required for war-fighting effectiveness. These warfighting leadership standards could then serve to focus elements of the leader development system--selection, training, assignment experiences, assessment, feedback, etc.--on preparedness.

Programmatic research on leadership performance and measurement at the Army Research Institute (ARI) supports the need to establish leadership performance standards. One program thrust has concentrated on simulated combat settings with objectives of (1) developing methods for measuring leadership performance in such settings and (2) applying these measures to establish the importance of leadership and components of it to unit performance. This paper describes trends obtained from this research with respect to leadership measurement and the relationship between leadership and unit combat effectiveness.

RESEARCH APPROACH

A test-fix-test approach has been used. Methods have been designed to measure components of leadership represented in U.S. Army doctrine on leadership. The measures have then been applied in simulated combat exercises to describe leader performance and determine relationships between leader and unit performance. Results have also been used to revise and, as possible, improve the measurement methods.

Use of doctrinal concepts of leadership has served at least two research functions. First, doctrine describes performances, operations, conditions, etc., which experience indicates are critical for success. As such, doctrine is a model from which expected relationships between phenomena (variables) can be derived. Thus, evidence supporting expected relationships, such as a positive relationship between leadership performance and unit effectiveness, is also indicative of measurement validity. Second, doctrine-based measurements result in a data base for decisions about the role of doctrinal concepts of leadership in combat settings.

¹The views expressed in this paper are solely those of the authors and do not reflect the views of the Dept. of the Army or the U. S. Army Research Institute.

The data collection setting--training exercises at the National Training Center (NTC)--has also been appropriate for inferences about leadership and unit combat effectiveness. For a training session (rotation) at NTC, two battalion task forces deploy to NTC and fight free-play, force-on-force (FOF) missions against a resident opposing force schooled in Warsaw Pact tactics. Multiple Integrated Laser Engagement Systems (MILES)--a simulation technology in which laser bursts are fired with blank rounds and detected by receptors on targets--is used on individuals and weapons systems to simulate and record firing data and kills. Resident observers/controllers (OCs) are attached to each unit to provide feedback and ensure that engagement rules are followed. Other, live-fire missions are conducted on terrain with pop-up and moving targets. The length, realism, and size of exercises make NTC an excellent setting for training as well as a valuable setting for research on doctrinal issues.

Two cycles of methods development and testing have been conducted. Both have involved observation and judgment of the leadership performance of platoon leaders (PLs) and platoon sergeants (PSGs) in NTC exercises. In Test 1 (see Rachford, Twohig, & Zimmerman, 1986), 10 broad leadership dimensions defined the framework for measurement of the performances of individual PLs and PSGs: planning, communication, supervision, teaching/counseling, technical proficiency, professional ethics, decision making, initiative, subordinate leader development, and soldier-team/cohesion development. Subject-matter-experts (SMEs) were trained in the measurement methods and, for purposes of data collection, accompanied platoon OCs during FOF missions. The SMEs made two sets of observations and judgments of PL and PSG leadership: (1) their performance of 90 relatively specific and objective leadership tasks (behaviors) during each mission and (2), after having observed several missions, summary judgments of PL and PSG performance on the 10 leadership dimensions. At the end of the rotation, the OC for a platoon also made summary judgments of the overall leadership effectiveness of the PL and PSG and of the effectiveness of the platoon with respect to mission accomplishment.

As Test 1 involved only one rotation, results had several limitations. Results nevertheless suggested the possible meaningfulness of the measures and identified methodological changes necessary for leadership measurement in dynamic and continuous field training exercises.

Test 2 (Twohig & Tremble, in preparation) built on the earlier results and changed the methodology for use by OCs (as opposed to extra data collectors) in NTC-type exercises. Similar to Test 1, methods for Test 2 produced judgments of leadership performance both for separate missions and for the rotation as a whole (that is, summary judgments). For each mission, platoon OCs observed their PL and PSG and made judgments of PL and PSG performance on five components of leadership: planning, supervision, communication, soldier-team development, and initiative. These components were selected to represent CAL's current and broader list of required warfighting leadership requirements and to fit with expected observation opportunities and workloads of OCs during NTC type exercises. For each mission, OCs also judged the overall effectiveness of their PL and PSG as a leader and the overall effectiveness of the platoon as a unit in accomplishing its mission. At the end of the rotation, OCs made summary judgments of platoon (unit) mission effectiveness and of PL and PSG performance on 11 leadership components: the five judged per mission plus

decision making, teaching/counseling, motivating others, trust in subordinates, flexibility, and technical/tactical proficiency. Test 2 has included data collection for three rotations at the NTC. Analysis of the OC judgments has been completed for the first of the three.

Two types of scores have been used in data analyses. One type consisted of scores for the summary (SUM) judgments made at the end of a rotation. These included: overall judgment of platoon/unit mission performance in the rotation by OCs (SUM unit performance), OC overall judgment of the leadership effectiveness of PLs and PSGs in Test 1 (SUM leadership effectiveness), and judgments of performance on the leadership components by PLs and PSGs (made by SMEs in Test 1 and OCs in Test 2). Judgments of performance on the leadership components have also been averaged as a SUM measure of leader performance (SUM leadership performance). The second type of score was a mission performance average (MPA) score, computed as the mean of the judgments of performance for the separate missions in the rotation. Accordingly, MPA scores were computed in Test 2 for overall effectiveness of the PL and PSG as a leader (MPA leadership effectiveness) and overall effectiveness of the platoon as a unit (MPA unit performance).

In Test 1, OCs and SMEs provided judgments for 24 and 15 platoons, respectively, with overlapping data for 12 platoons for correlations between SUM leadership performance (SME based) and SUM unit performance (OC based). For Test 2, MPA scores could be calculated for 24 platoons, and SUM scores were available for 31 platoons, with samples of 20-24 for intercorrelations.

Consistencies in results of the two tests are discussed below. Test 1 results were originally and more completely described by Rachford, et al., (1986). More complete reporting of Test 2 results is in preparation (Twhig & Tremble, in preparation).

RESULTS AND DISCUSSION

There has been consistency in judgments of leader performance. This consistency has been found between different observers of the same platoons (Test 1) and within observers of the same platoon at different times (Test 2). As Table 1 shows for Test 1, SUM judgments of leadership performance by SMEs and of leadership effectiveness by OCs were significantly correlated for the PSG. For Test 2 (Table 2), MPA judgments of leadership effectiveness were positively correlated with SUM leadership performance for both the PL and PSG. These results are indicative of concurrent (Test 1) and discriminant (Test 2) validity. However, halo effects are apparent, and more research is needed to confirm.

Perceived quality of leadership has been positively related to perceived unit performance. Tables 1 and 2 also show correlations between judgments of leadership and unit performance. The positive correlations further suggest measurement validity since doctrine and the scientific literature support a positive leader-unit performance relationship. There is a need to confirm these results using more independent measures of leader and unit performance. Current research on measurement of unit performance (e.g., Root, 1987) should enable this.

One research objective is to assemble a data base for establishing the importance of leadership and the various leadership components to unit performance. Both tests have not only provided evidence that SUM leadership performance is positively related to unit mission performance. They have also produced some possible leads about the relative importance of the leadership components examined. For example, three of the five leadership components with the relatively strongest correlations with unit performance were the same in both tests (Test 1, $r = .5$ to $.72$; Test 2, $r = .67$ to $.75$): planning, teaching/counseling, and soldier-team development. The remaining components were technical proficiency ($r = .63$) and ethics ($r = .54$) for Test 1 and supervision ($r = .73$) and decision making ($r = .80$) for Test 2.

PL-unit performance correlations have been higher than PSG-unit correlations for SUM measures of unit performance. As both Tables 1 and 2 indicate, SUM judgments of unit performance tended to be more strongly correlated with the perceived leadership of the PL than the PSG. This appeared for all SUM judgments of leadership in Tests 1 and 2 and for MPA leadership effectiveness in Test 2.

The higher PL-unit performance correlation is consistent with conventional role expectations of the PL as the main director of activities in the field. Since the PSG is expected to have relatively more operational control in garrison, a stronger PSG-unit performance correlation might be predicted for leadership in the home station environment. The latter prediction was supported by home station data collected along with the Test 1 unit performance data (Twohig, Tremble, Rachford & Williams, 1987).

As Table 2 also shows, a relatively higher PL-unit performance association was not obtained for the MPA measures of leadership and unit performance. This difference in patterns for the MPA and SUM measures may suggest random variation. The consistency of the SUM pattern in both Tests 1 and 2 suggests otherwise. It is likely, for example, that MPA and SUM measures have somewhat different meanings: MPA values are bounded by the leadership manifested in separate missions while SUM judgments allow for retrospective integration of performance across missions and, possibly, performances outside the combat missions assigned to the units.

Evidence has suggested the importance to unit performance of the PL-PSG as a leadership team. In addition to leadership and unit performance, both tests have captured judgments on the relationship between the PL and PSG. Judgments of PL-PSG role clarity have been consistently correlated unit performance. Tables 1 and 2 also suggest the importance of the PL-PSG relationship in that platoons with effective PLs tended to have had effective PSGs as well.

Table 1

Summary Correlations of Leader-Unit Performance by Different Observers In
Test I: Observer Controllers (OCs) and Subject Matter Experts (SMEs)

	<u>SME</u>		<u>OC</u>		
	<u>PL</u>	<u>PSG</u>	<u>PL</u>	<u>PSG</u>	<u>UNIT</u>
<u>SME</u>					
PL	X	.80	.49	.57*	.55*
PSG		X	.30	.66*	.43
<u>OC</u>					
PL			X	.70*	.80*
PSG				X	.66*

Note: * $p < .05$; Sample: 24 platoons for OC-OC correlations;
15 platoons for SME-SME; 12 platoons for SME-OC.

Table 2

Overall Leader and Unit Performance Correlations in Test 2:
Mission Performance Average (MPA) and Summary (SUM)

	<u>MPA</u>			<u>SUM</u>		
	<u>PL</u>	<u>PSG</u>	<u>UNIT</u>	<u>PL</u>	<u>PSG</u>	<u>UNIT</u>
<u>MPA:</u>						
PL	X	.60*	.65*	.79*	.53*	.77*
PSG		X	.66*	.30	.75*	.42*
UNIT			X	.52*	.64*	.74*
<u>SUM:</u>						
PL				X	.69*	.80*
PSG					X	.66*

Note: * All correlations .42 and above $p < .05$.
Sample: 31 platoons for SUM-SUM, 20-24 for others.

SUMMARY

The consistency of results for Tests 1 and 2 suggest the potential for development of doctrinal based measures of leadership for use in a field training environment. Development of leadership performance measure faces certain difficulties. Some of the difficulties have been identified in interviews and field notes on use of the methods in Test 2. These include: (a) variations in a unit's (and its leaders') involvement in the various phases of a tactical operation (e.g., planning, preparation, execution), (b) limited behavioral discretion by the leader/unit targeted for observation due to the actions/decisions of superiors, and (c) possible irregular opportunities for observation of key leader actions or the conditions pertinent to them.

Future research in this program includes analysis of data from the other NTC rotations in Test 2 to confirm and add to current trends. Those data will also be informative of measurement reliability since company OCs as well as platoon OCs made performance judgments of platoon leadership and unit performance. Data collection is also planned for other training environments to identify the generality of present approaches and findings.

REFERENCES

- Rachford, D. L., Twohig, P. T. & Zimmerman, R. A. (1986). The platoon leader - platoon sergeant in a tactical environment: Focused rotation at the National Training Center (NTC). Working Paper, LMTA 86-06, U. S. Army Research Institute.
- Root, J. T. (1987). A unit effectiveness measurement system. Proceedings of the 29th Annual Conference of the Military testing Association, 445-450.
- Twohig, P. T., Tremble, T. R., Rachford, D. L., & Williams, R. (1987). Platoon climate and platoon performance. Proceedings of the Third Annual Leadership Research Conference, Volume 1, 181-191. Center for Army Leadership, Fort Leavenworth, KS.
- Twohig, P. T., & Tremble, T. R. (in preparation). The development of leader performance measures in a tactical environment: preliminary results. Research Report, U. S. Army Research Institute.

Attitudes of GFAF Reservists Towards Reserve Duty Training

Sibylle B. Schambach
Federal Office of Defense Administration, Bonn, FRG

General statements

Within NATO's strategic concept, the German Federal Armed Forces (GFAF) provide for a constantly present manpower of 495,000 active soldiers which may quickly grow to a strength of 1.34 million soldiers in the case of tension or war.

To meet these requirements, the GFAF keeps

- 250,000 regular and temporary soldiers
- 206,000 conscripts during regular duty
- and a strong reservist component.

Reservists are trained soldiers back in civilian life but in different states of availability for activation. In peace, they contribute their share to the high level of operational readiness of the in-place forces of the "Bundeswehr" by serving for reserve duty training periods. In the case of mobilization, reservists form the largest personnel component, enabling the forces to expand rapidly to full strength and effectiveness.

Declining birth rates will take effect on the GFAF as well as on other fields of trade and commerce where young men are wanted. The strength of the age groups for conscription will be cut by half in the early nineties. One of the measures taken by the Federal Government to cope with the problem, is to prolong the regular duty period of conscripts to 18 months (from 1989 on). Another measure is to increase essentially the reservist component in manpower level and operational readiness.

The number of openings for reserve duty training is to be increased from 5,000 in 1983 to about 10,000 in 1992, a training period lasting 11 or 12 days on average. 10,000 openings mean that up to 350,000 reservists will have to be trained during that year. The number of call-up procedures to be dealt with by the administration, will even be higher: In 1987, for 6,600 openings 320,000 reservists had to be called up or contacted of whom 218,000 finally performed training, that is 68 %. The remaining 32 % dropped out mainly from medical reasons and on grounds of "individual hardship" (these were, in most cases, a current study or vocational training). Only some 4 % of the called-up dropped out from occupational deferment.

The conclusion to be drawn from these facts is that the number of openings for reserve duty training cannot be increased without limitation. Evidence should be gained whether the high rate of drop-outs might be connected with a small readiness of reservists to render reserve duty training and how their readiness might be enhanced.

Reservists' Opinions On and Attitudes Towards Reserve Duty Training

A series of investigations were carried out in order to answer the question asked above, with the following subgroups:

- reservists
- experts of GFAF call-up administration
- population in general
- "opinion leaders"

I will mainly refer to a questionnaire study carried out by the Federal Armed Forces Office in 1987 with a representative sample of 3,000 GFAF reservists on their personal readiness for training.

The sample was so constructed that one third of the reservists were questioned at the beginning of their training period, another third of them at its end; the remainder were not in training. 58 % of the questionnaires were returned so that 1,758 reservists finally took part in the investigation.

The reservists were asked to choose how often they would be ready to render training periods:

Frequency Preferred for Repeated Retraining Periods

- N = 1758 reservists -

one retraining period every year	7,8 %
one retraining period every two years (current practice)	45,3 %
short retraining periods on weekends	2,6 %
combined procedures	9,6 %
none of the above	34,8 %

The table shows that about 45 % of the respondents would rather cling to the current practice of one retraining period every two years. Reserve training once a year was favoured only by 8 % of the respondents. Training periods at even shorter intervals underly certain restrictions for conscripts after their regular duty, the aim being to let these young men first follow their civilian career for some time before they are called up again.

Remark that reserve duty training on weekends does hardly find any support. Details on this fact will be given further down.

Almost 35 % of respondents, when inquired about frequency of reserve duty training, did not support any of the regulations proposed. Many suggested to "abolish reserve duty training".

For what reasons do reservists reject an increase in training periods? Intelligence on this question may be drawn from certain

indicators: For example, other surveys show that the attitudes of soldiers towards military duty will correlate highly with the attitudes these soldiers perceive in their personal environment. The present study therefore investigated into the reactions that reservists expect among their families, friends, and occupational surroundings, on an increase in reserve duty training.

Reactions of the Social Environment
on More Reserve Duty Training

- opinions of N = 1758 reservists -

reaction of	reactions			
	positive	neutral	negative	don't know
father	10,1 %	46,9 %	30,9 %	12,2 %
mother	6,2 %	50,1 %	35,4 %	8,2 %
female partner	3,3 %	21,9 %	69,7 %	5,2 %
friends	3,4 %	47,1 %	43,3 %	6,1 %
employer	2,8 %	23,9 %	73,3 %	-, -

As can be seen from the table, reservists ascribe highly negative reactions to their female partners. Accordingly, another question gives 79 % of respondents who feel that more reserve duty training might be detrimental to family life. The above-mentioned finding that weekend duty is not favoured by reservists, may spring from figuring it to impair family life. Moreover, the reactions of spouses and female partners seem to be prevalent, in this question, upon the reactions expected in employers who should prefer employees to render reserve duty training beyond working hours.

An even more negative reaction on an increase in reserve duty training is ascribed to employers of reservists. Further inquirement confirms that 77 % of respondents will in this case expect detrimental effects on their occupational careers. This estimate is supported by results of another investigation carried out among "opinion leaders", many of whom are in the position of employers:

According to opinion leaders, about two out of three employers (69 %) will not like their employees to render more reserve duty training but would tolerate them to do so. The same will not be valid, though, for reservists who belong to an establishment's management. Regarding them, the opinion leaders expect that every second employer would

impose on them to let themselves be discharged from military duties, or the employer himself would apply for their occupational deferment.

44 % of opinion leaders even hold it possible that employers might rather engage applicants who are not subject to reserve duty training.

Reservists therefore do not seem unrealistic in their judgment of employers' behavior.

From the indicators described above it may be derived that the attitudes of reservists towards more reserve duty training are mainly determined by a cost-to-use deliberation. Of crucial importance is to what degree reserve duty training will run counter to personal interests.

These attitudes have to be taken into account when reservists report their experience regarding reserve duty training:

Reservists were first questioned about their contact with the authorities in charge of their call-up. Some 25 % of respondents report negative experience with the administration. They find their individual concerns - especially occupational and other personal demands - having been disregarded. Nearly 40 % of the reservists feel they were not sufficiently informed on the purpose and course of their training period, which was much disapproved.

Reservists were then asked to report on their experience during the training period itself.

Own Experience During Reserve
Duty Training

with respect to	all in all positive	positive with reservations	not positive
military leaders	43,8 %	44,0 %	12,2 %
comrades	57,6 %	36,0 %	6,4 %
organization of military service	12,5 %	41,0 %	46,5 %

Positive reports are prevailing with regard to superiors and colleagues, whereas almost 47 % of respondents criticize the way the duty period was organized. Disapproval was highest among the subgroup of reservists who had just started training. Encouraged to give further details in their own words, reservists mentioned:

Negative Experience During Reserve Duty
Training

- numbers of responses -

unreasonable conditions of transport, lodging, material, medical care, stress, etc.	306
lack of organization	181
lack of reasonableness in general	101
arrogance of military leaders	80

In spite of these negative judgments, 58 % of the reservists maintain that during their time of active duty they had liked, or rather liked, to be a soldier. Even, more than 75 % of respondents indicate their general attitude towards the Bundeswehr to be positive or neutral and only 8 % say it is negative. Demographic characteristics of the sample questioned give hints why reserve duty training is all the same so badly estimated among the reservists:

Nearly 80 % of the respondents were, at the time of the investigation, aged between 22 and 28 years; just about 2 % were younger than 22.

Most of them (over 75 %) were engaged or married.

84 % of respondents had already finished vocational training or obtained a university degree.

96 % were employed.

To conclude, a large percentage of these men have engaged in long-term personal relationships and are in the middle of their occupational careers. Very often, their military ranks and assignments do not equal the responsibilities attached to them in civilian life. Also, payment granted for reserve duty training follows the rates for young recruits during regular duty and will not compensate for the losses of persons who run their own

business. Consequently,

- * reservists feel ill-treated when merely addressed in terms of command and obedience. They want to be individually respected and sufficiently informed; they expect training to be well organized, sensible, and useful so that their individual burden is worth while;
- * reservists demand "just payment" and the return to their civilian job legally guaranteed;
- * reservists would like service terms be coordinated with personal demands and service to be near home;
- * reservists feel misplaced if their assignment does not correspond to their occupational or military training.

The conditions to make such feelings and demands arise, are set up by the Bundeswehr itself and can therefore be handled and changed in order to improve the individual readiness for training. When subgroups were compared, respondents who had indicated positive attitudes towards the Bundeswehr were found to be most susceptible of improval of training conditions.

Summary

Owing to rapidly declining birth-rates, the share of conscripts in the Geman Federal Armed Forces (GFAF) will go down severely from the early nineties on. One of the measures taken to maintain the prescribed personnel strength, is to enhance essentially the reservist component. By multiplying the openings for reserve duty training, more reservists shall more frequently be called up for retraining periods.

By contrast to these requirements it is observed that the readiness of reservists to participate repeatedly in training, is comparatively small. Empirical studies show that, at present, an individual burden higher than one retraining period every two years would not be understood and therefore not be supported. Reservists' views of repeated retraining are found to depend not only on attitudes towards the "Bundeswehr" and on estimates of their social surroundings, but just as much on factors linked to organization, administration, assignment, behavior of leaders, personal experience, and amount of refunds during duty periods.

It is therefore suggested that the acceptance of additional reserve training be increased by harmonizing the call-up of the individual reservist with his personal needs and situation. Also, leader behavior in reserve units should be improved in order to fit the respective age-groups.

ACTUAL CONTRIBUTIONS OF MILITARY PSYCHOLOGY TO MANPOWER MANAGEMENT

Friedrich W. Steege
Psychological Service of the GFAF
Federal Ministry of Defense, Bonn, FRG

Introduction

My paper is hermeneutic and policy oriented rather than experimental in nature. It is my conviction that military establishments need decision aiding reflections as much as experimentally proven data and inferences.

Manpower management in the German Federal Armed Forces (GFAF) is strongly influenced by societal factors. This is especially true with respect to draftees and reserve forces, but is also valid for medium- and long-term personnel. That is the reason why I will give at first some details on recent insight and public opinion results concerning the political culture and security policy in my country.

Several inquiries have been conducted recently by the Psychological Service of the GFAF to get a clearer picture as to the measures to be taken in manpower management. Thus, I will report here in a little more detail on two studies, (i) a survey of the job satisfaction of volunteers in the ranks of men and of non-commissioned officers respectively, and (ii) a survey concerning the future number and quality of career officers. There will be other German contributions to this conference pertaining to general manpower management questions. They cover (i) main results of a comprehensive survey concerning reserve forces (Schambach, 1988), (ii) the recent development of computer assisted testing (Wildgrube, 1988) and the functionality of our new CAT system (Habon, 1988), and (iii) some improvements in our personnel selection strategies (Kaiser, 1988; Melter, 1988).

In the final part of this paper I will present some new measures designed to improve personnel management for medium- and for long-term personnel, especially for career officers.

Societal Factors as Reflected in Recent Investigations

In addition to the well known fact of declining birth rates, the GFAF, and comparably other European allied forces, face novel wide-spread legitimation problems. Many of the young people as well as the society at large are more and more reluctant to perceive as predominant the threat in the military area. They regard as far more important the global survival as threatened by a destroyed environment, and in some cases the health hazards caused by modern epidemics. Especially since the fruitful and hope-inspiring talks, and the INF-agreement, the United States and the Soviet Union have achieved in the disarmament arena, the awareness of a military threat is vanishing. This does affect, of course, the readiness of our youth to render military duty.

What are the current figures reflecting related attitudes of young German men? Each year since 1979, a civil public opinion institute has been conducting a sample inquiry. With respect to the general attitude towards military duty, the following figures were obtained: In 1984 48% of the young men questioned saw a military threat as compared with 33% who did not feel so (19% didn't know). In 1988, this figure has reversed (20% see a threat, 62% don't). That has led to the

rather surprising following answers: In 1984, the USA was considered to have a greater interest in disarmament (26%, as compared with 7% by the USSR), in 1988 the figures have also reversed (USA: 11%; USSR: 30%). The growing hope with respect to peaceful politics is connected with increasing reservations against the military in general and against the use of nuclear weapons in particular: In 1984, 20% of the young men expressed a positive general attitude towards the military, in 1988 only 12%. A negative general attitude was shown by 31% of the inquired in 1984, and by 48% in 1988.

Of special importance are some current conditions affecting young people growing up in the Federal Republic as put together by social psychological and developmental research:

- ° The possibilities of the up-growing generation of shaping their own life have shrunk.
- ° The length of time spent in educational institutions has increased considerably: A rising number of young people remain economically dependent until they reach a relatively old age (the unemployed, the students, members of the so called alternative scene), whereas they are counted as adults in a psycho-social sense (this period is called "post-adolescence").
- ° In the same time, there are more intensive claims for self-reliance, autonomy, and independence, claims that are furthered by the market dependent consumption, and by leisure-time industry.
- ° The potential influence of adults in family, school and professional life is decreasing, instead partner-oriented communication is preferred. Psychoanalytically there are signs of a new type of ego-weakness. Some researchers see a corresponding danger of egocentrism rooting in too high stress on the individual.
- ° Furthermore, there are clear signs of crisis, even though they cannot be quantified easily. Expressions and types of deviant behavior have to be taken very seriously, particularly the self-destructing acts that are often interpreted as helpless trials of escape.
- ° Individualization, finding of one's own identity, self-realization, and internal distance from traditional institutions like church, matrimony, family, and profession lead more than heretofore to sub-cultural behavior and groups.

In summary one can say that youth today are confronted with a comprehensive process of modernization and, in the sense of Max Weber, total rationalization of society and culture. Opposed to that they experience enormous horizons of expectation and desire. As a consequence the number of those young people who deliberately "disembark" from mainstream culture has increased. The majority of young people are confronted with problems that are not their own ("youth") problems.

As far as armed forces are concerned, the finding is important that young men feel more distance from traditional institutions and at the same time question more intensively conditions and meaning of their military service. If they do not question military service in principle, they are at least looking for opportunities comparable to civil and/or professional life. This is the topic of the following paragraphs.

Satisfaction of Enlisted Soldiers with their Initial Assignment

Only a few days ago, I received first results of an inquiry of enlisted men who were assigned to special military jobs and received a higher entry pay than regular enlisted men (Rank: Private 1st Class). We have a regulation that gives them the chance of serving in a preliminary status and of deciding after four months of service whether they want to enlist for a longer term or serve as a draftee.

About 1,500 soldiers received a questionnaire by mail, about 1,000 of them responded. In this context, I will only report some data with respect to (i) the reasons why they enlisted in this special category, (ii) the satisfaction scores the soldiers showed, and (iii) the success of this measure.

Most of the soldiers regarded the preliminary status as a chance of testing their possible longer-term enlistment, and of getting to know the Bundeswehr. They were satisfied with their military occupational specialty that fitted and improved their professional training, and with the possibility to choose their location of duty freely. They also praised their opportunity to have a responsible task/position. Of lesser importance seemed to them a possible unemployment or dissatisfaction with their civil workplace.

The satisfaction the soldiers showed was mainly related to location and position (military occupational specialty). Only 10 % of the sample were dissatisfied with both. The two variables "satisfied with location" and "satisfied with present job" are highly correlated ($p < .001$) with the criterion variable (readiness to keep enlisted). The analysis of the frequency distribution shows that primarily those soldiers not satisfied with their locations (36 %) or with their jobs (51 %) deny further enlistment to a much higher extent than is to be expected in a random setting.

In general, those soldiers who do not want to keep enlisted rate every aspect of a military job asked for more critically or negatively. Many of them complain at not having a job assigned to them which provides them with the opportunity to apply their professional knowledge and skills. Hence, they regard their service as little effective. The great majority, however, is willing to prolong their terms. Consequently, this measure is very successful in an overall assessment.

Assignment of Career Officers

In late 1987, the Psychological Service has been tasked to investigate current problems concerning the number and the quality of career officers in our GFAF in the years to come. The aim of the survey was to get additional evidence with a view to decisions to be taken to secure a sufficient number of temporary-career officers for a commission in the medium term.

The GFAF need some 650 career officer candidates per age group for line duty in order to maintain operational readiness and achieve a balanced age structure. Only one third of them are enrolled from the outset as career officer candidates. Most officers must qualify for commissioned status by their performance in the armed forces, serving first as temporary-career personnel.

We are currently facing discrepancies caused by the requirements of our personnel management and the special aspirations of officers having graduated from the Bundeswehr universities. This is especially true with respect to technical engineering and computer science courses of study. There are indications that in these faculties the majority of students regard their diplomas as a prerequisite for a job in private industry rather than for a career as officer. This may cause difficulties in satisfying the *Bundeswehr* requirements for career officers in the medium term.

The following *reasons* causing the problem situation have been determined: (i) "Workplace" *Bundeswehr*: i.e. discrepancy between university courses of study and reality of the job (jobs differ considerably from contents of study courses); discrepancy between wage expectation after graduation and income offered by armed forces; relatively low job satisfaction of graduated officers. (ii) Personnel Management: Negative effects emanating from the alleged seniority principle; negative repercussions of the assignment backlog; requirements of mobility. (iii) Societal Factors: Family situation; excessive duty hours as dictated by the mission; Little attractiveness of many garrisons.

These reasons are basically determined by at least three *conflicting goal areas*: (i) Personnel management aspects vs. self-comprehension of the armed forces universities. The personnel management of the armed forces is directed by the necessities of military assignments. The self-comprehension of the armed forces universities is primarily shaped by the comparability with scientific education at civilian universities. I.e., they view first of all their genuine (academic) role. (ii) Effectiveness vs. attractiveness of an officer career in the GFAF. Positions that require graduation are expected to guarantee higher prestige and better advancement possibility. If a temporary-career officer feels misled in these expectations, he may tend to leave the armed forces and not seek a commission. This may be particularly true if private industry is trying to offer contracts long before the end of an officer's temporary term. (iii) Military ethics vs. integration into society. Officer candidates tend to adopt the wrong attitude that courses of study are the main element of their profession during their long years on the university campus. Military leadership has on the other hand to further the attitude that courses of study are part of the officer training. The principal aim of this training is to prepare the officer for his profession, i.e. for his military duty.

In December 1987, the questionnaire was mailed to about 4,200 officers or officer candidates (students at the armed forces Universities) of those five age groups that have commenced studies in 1981, 1982, 1983, 1985, and 1987. We included officers/officer candidates (i) after having entered an AF university, (ii) after having received the first university degree, (iii) after graduation, and (iv) after one to two years following their return to military life. 67 % of the officer candidates/officers returned the questionnaire spontaneously.

Main Results of this investigation were: The majority of the inquired officers regard their profession/job as a responsible life-time task. Positively evaluated are primarily the variety, alternation and the broad horizon as well as the security of the job (work place) and the secured future. As negative are mainly expressed: the mobility requirements, the excessive duty hours (many units serve for more than 56 hours a week) in connection with irregular duty times, and for many of them a lack in transparency of personnel management/planning. Increasingly, the family members regard the mobility requirements as a burden, especially with respect to the job of the spouse and schooling of the children. In summary, the results of the investigation confirmed the main reasons determined earlier.

An analysis of the results reported to the Federal Minister of Defense has identified three categories of measures to be promoted: (i) Short-term measures: The mode of applying for, or being taken over as, a career officer will be altered, the approval may be given earlier than before and more flexibly in terms of time (up to now only once a year). In keeping valid the general rule that officer candidates complete courses of study, more frequent exemptions may be made as caused by the given situation: Increased number of officers taken over for a regular career although they didn't successfully finish their courses of study, increased number of temporary-career officers accepted without courses of study after they have demonstrated their aptitude as troop commanders over a sufficient period of time. Parallel to these first measures, the quality of information to be provided to the possible applicant will be improved. Commanding officers are urged to motivate young officers and to explain career perspectives in greater detail than formerly. (ii) In the medium-term, the following measures shall be realized: Scrutiny of those positions that presuppose university courses of study as an indispensable requirement; scrutiny of the demand of career officers in technical engineering and computer science positions; if necessary, increase of the number of officers to be taken over from ranks; if necessary, increase of the rate of young officer specialists to be taken over; utilization of the reservist organization to advertise re-employment of former career officers. (iii) General supporting measures: Most prominent among the recommendations given are the regulation compensating for additional duty hours, a new quality of personnel management that gives more regard to the individual and that takes better

into account family matters (see next paragraph), the reduction of the number of transfers, and a payment that is adjusted to academic degree, responsibility, and performance.

A New Quality of Personnel Management

Armed forces must not wait to be met by society, they have to integrate into society as far as possible in view of their mission. With this basic rule in mind, our Personnel Division is looking for new ways to fulfil its goals.

Among the proposals derived from the results of analyses and investigations reported above for career officers are: (i) Longer assignment periods and greater transparency of assignment planning; (ii) Involvement of families in personnel decisions; (iii) A more efficient promotion policy, particularly with a view to duty performance.

The FMOD, especially the Chiefs of Staff council, is presently coordinating new measures to improve personnel management. After a new Chief of the Personnel Division has assumed office in April 1988, these measures have been promoted and discussed. Consensus has been reached as follows:

At present, personnel management is still extraordinarily successful. The highest strength figures of career soldiers and temporary-career soldiers have been achieved since 1956. At the same time, however, it has become increasingly difficult for the soldiers to strike a balance between duty requirements and personal aspirations. The first and foremost aim of the new placement strategy is therefore to rely on a recently improved long-term individual assignment planning, so that a stable individual perspective covering the present and the consecutive positions can be established. For the soldier this includes greater individual involvement and the possibility to plan his private living conditions, for the Personnel Division this means greater self-constraint and for the operating units short-term vacancies cannot be excluded.

In more detail we are planning as follows: (i) Greater *reliability* of assignment planning. In future, for every officer an individual assignment schedule will reflect the length of time to be spent by him in a job. Personnel in commanding officer positions will stay for three years at least. The last assignment before retirement will be determined in a personal interview at the latest five years before an individual reaches his retirement age. Any deviation from the pre-determined tour of duty will not be allowed unless the soldier expressly agrees. (ii) Greater *transparency* as to his further assignment. In future, every officer will be notified on his further assignment prior to the end of his training, and at the stage of his first advancement rating. This may be the confirmation that he will remain in the same career, or the announcement of a change. (iii) The *participation* of the soldier in personal interviews will be improved. The soldier may, if he wants to do so, bring his spouse, or other relatives, to the interview. (iv) *Mobility* in the Armed Forces. Generally, the principle that a soldier must accept a transfer at any time is still valid. Where possible, however, a regional assignment and promotion planning is intended. Moreover, the individual soldier may eventually decide, whether or not his family will move to the new duty station. Nevertheless, officers, especially those having higher command responsibility, are being urged to have their families move to their place of duty. (v) Personnel management and armed forces organization. It is intended to more intensively *communicate in advance organizational changes* to personnel management so that personnel measures can be adjusted to the changes in a timely fashion.

Conclusion

As you have seen there are many problem areas with respect to society we have to take into account in our respective personnel managements. The motives of young men in the Federal Republic for initial assignment reflect a close connection with individual life planning, professional

training, and preference to serve at the location closest to the individual wishes. Not so obviously but pointing in the same direction, career officers plan their service and express their problems. Personnel management is well advised to seek a balance between societal requirements and genuine military aims.

Thus, armed forces are - in contrast to many arguments used in the past - affected by, and integrated into, societal, cultural, and more and more *global* developments. The most recent mutual visits of highest ranking defense officials and military of East and West may point to a beginning new awareness as to this kind of responsibility for global survival. As psychologists we have to further emotional processes to the point that members of military organizations are able to keep up with these signs of hope and at the same time don't forget their mission.

Reference Notes (Papers presented during the 30th MTA Convention, Arlington, VA.)

- Habon, M.W. (1988). *Functionality and architecture of the GFAP computerized adaptive testing system.*
Kaiser, A.H. (1988). *New assessment for short-service volunteers.*
Melter, A.H. (1988). *Methodical and organizational development of selection: Progress and results.*
Schambach, S.B. (1988). *Attitudes of GFAP reservists towards reserve duty training.*
Wildgrube, W. (1988). *Computerized testing (CAT) in the GFAP.*

LEADERSHIP ISSUES IN GENDER INTEGRATION

Major R.A.V. Dickenson

Royal Military College of Canada
Kingston, Ontario, Canada

Background

Expanded mixed-gender employment in the Canadian Forces (CF) is reflective of changing trends within society. These include the expansion of women's rights and obligations, an increased participation of women in the labour force, the increased movement of women into previously male-dominated occupations and vocations, lower birth rates and delayed marriages, the shrinking labour pool of 17-24 year olds, and the evident desire of many women for the equal opportunity to participate with and compete against men in the world of work. In Canada, equal status and employment for women have been the subject of government recommendations (e.g., Royal Commission on the Status of Women, 1970), legislation (e.g., Canadian Human Rights Act, 1978), and specific bases of discrimination, including gender, have been proscribed in law by the Charter of Rights and Freedoms (1982). Clarification in the form of recommendations was also more recently provided through a government committee report "Equality for All" (1986).

By 1974, the CF had adopted all Royal Commission recommendations except those which would open all military occupations to women. Ensuing studies (e.g., Servicewomen in Non-Traditional Environments and Roles (SWINTER), 1980-85), policy reviews, and directions from the Chief of Defence Staff and Minister of National Defence have resulted in further expansion of mixed-gender employment opportunities. For example, by 1986, expanded employment opportunities for women included a number of previously all-male units (e.g., Army combat service support and Navy supply ships) and the majority of previously all-male occupations. In addition, trials were initiated in 1987 to evaluate any effects of mixed-gender employment on the operational effectiveness of combat units (Combat Related Employment of Women (CREW)). In these trials women are being trained and will be employed in the remaining (combat) occupations, (e.g., Artillery (ARTY), Maritime Surface and Sub-surface (MARS)).

Gender Integration as Change

Mixed gender employment in the CF has been approached from within a context of organizational change. Historically, the military has been male-dominated, with deeply embedded attitudes and behaviours. Introduction of women into previously male military roles would be expected to be problematic, and into combat support and combat roles, even moreso. Therefore, integration of women into the military can be seen as a special type of change that goes against the values and attitudes of many males. For this change to be successful, it must be carefully managed.

The views and opinions expressed in this paper are those of the author and not necessarily those of the Department of National Defence.

Lamerson (1987) outlined a model for gender integration, indicating that:

- a. acceptance of servicewomen by servicemen is a necessary requirement for meeting unit objectives and maintaining operational effectiveness;
- b. for the entering groups to be successfully integrated into the existing groups, the former must be recognized as contributing equally to unit objectives; and,
- c. it is important to be informed on the reactions of the existing group, peers, subordinates, supervisors, and entering group.

Lamerson further noted that in such a change situation, the organization (CF) could:

- a. take no action;
- b. create favourable conditions in the environment;
- c. educate men on the abilities of women;
- d. prepare women to handle negative reactions, environmental traditions, and specific unit requirements; or,
- e. implement a combination of b, c, and d.

For successful integration, the organization should manage the change, and CF policies were developed to do just that, i.e.:

- a. prepare the existing group;
- b. prepare the entering group;
- c. prepare the environment;
- d. evaluate any special interventions; and,
- e. evaluate overall organizational impact.

Leadership Role

The Canadian Forces have adopted the definition of leadership as "the art of influencing human behaviour so as to accomplish a mission in the manner desired by the leader" (Canadian Forces Publication 131). While recognizing that "leadership is the most observed and least understood phenomena on earth" (Burns, 1978), it is also obvious that the nature of good leadership does not depend on whether or not a unit is mixed-gender. Sound principles of leadership can be effectively applied and adapted to any situation, including mixed-gender (change) situations. Examples of leadership principles include leading by example, training subordinates as a team and employing them up to their capabilities, and knowing subordinates (Department of National Defence, 1978).

Successful gender integration and mixed-gender employment is a leadership issue, where it is the responsibility of the leaders to ensure the change is managed effectively. Resch (1988) identified the following major lesson from the SWINTER trials:

The greatest threat to combat effectiveness, cohesion, and morale in mixed-gender units is the mind set of the leader and his subsequent overt and covert behaviour toward women in the military units.

Most of the failures of leadership arose through lack of information and preparation on the part of the leaders, not basic competence (Resch, 1988). What is critical for the leader and his/her subordinates in a mixed-gender situation is INFORMATION. The CF recognized that there has been so much misinformation and lack of information on gender integration and women, that a major education and information program was needed as a key part of any mixed-gender implementation program. Such programs have now been developed and are being implemented across the CF, through the co-operative actions of CF senior leadership and the operational commands.

Lessons from Past Experience

The programs are based on lessons from past experience, sound leadership principles, scientific information, and policy review. For example, the findings from over 50 behavioural science reports on the SWINTER trials, as well as information from operational reports and from other countries have been codified and incorporated into the programs. Observations from past experience include those involving attitude patterns, effects of rumours, focusing of resentment, stress, common distortions in leader perception, and effects of prejudice.

Empirical information from the fields of psychology, sociology, biology, physiology, etc. has been utilized to clarify and dispel myths concerning gender differences in ability and performance. Policy reviews were undertaken of such issues as fraternization, harassment, and pregnancy, which led to revisions and clarifications that are more consistent with the expanded mixed-gender employment situation. And finally, reassurances are built into the programs, to address and underline the following concerns of leaders:

- a. primacy of operational effectiveness;
- b. natural resistance to change;
- c. trust in the selection and training system; and,
- d. the fact that everyone must learn to deal with change.

Education Packages

The Land (Army) and Sea (Navy) operational elements have instituted education programs specific to their operational environments, based on education packages developed under contract by a former Social/Behavioural Science Advisor to the SWINTER Land Evaluation. The packages were designed to prepare officers and non-commissioned officers to:

- a. recognize gender-related challenges to discipline and morale which may occur in mixed-gender units; and,
- b. develop appropriate strategies to meet these challenges.

These packages were initially given by the Command Headquarters staff to formation, base and unit Commanders and staff, who in turn are using the information for the education and preparation of their own personnel.

Each package contains briefing notes, briefing texts, references, and slides. Each is tailored to the specific operational environment and the information is typically organized as follows:

- Background - The Canadian Experience
- Advantages of Mixed-Gender Groups
- Recorded Observations (Initial Integration)
- Effects of Prejudice
- Countering Prejudice
- Sex Differences
- Relationships
- Leadership - Individual Considerations
 - Group Considerations
 - Welfare & Morale Considerations
- Preparation of Servicewomen
- Spousal Concerns

The basic Land environment briefing is based on one package, "Mixed-Gender Service in Army Field Units". The first presentation was given by Command Headquarters staff (including two Personnel Selection Officers and a female Medical Officer) to the Commander and unit COs of 1 Canadian Brigade Group in January 1988. Guidelines and assistance were also given to the commanders for the subsequent preparation and implementation of their own information and education programs. All commanders and CCs have now received the education package/presentation and are responsible for educating their own personnel.

The Sea environment has developed three information packages:

- a. Mixed-Gender Service: Commanders' Considerations (for ships' Captains and Executive Officers);

b. Mixed-Gender Service: Chiefs and Petty Officers; and,

c. Mixed-Gender Leadership (a new chapter to the "Guide to the Divisional System", which is a handbook for junior officers and senior non-commissioned members).

All Captains, XO's and Coxwains of every ship are given the package. In addition, mixed-gender briefings are given to the PO2/Sgt professional development workshops, the Coxwains' course, and the CO/XO refresher course. Regular material updates are built into the packages based on any policy changes and additional relevant material, as published from the research and other countries.

Leader Training

Similar education packages are being developed for use within the CF Training System (e.g., Canadian Forces Officer Candidate School, Leadership Academy, Recruit School, etc.), to be incorporated into officer, junior and senior non-commissioned officer training, and recruit-instructor training. CF leaders also obtain information and education on leadership and gender-related issues by various other means. For example, all officer cadets at the Canadian Military Colleges (CMCs) are taught military leadership and applied psychology in all four years as part of their academic curricula. Topics such as values, attitudes, prejudice, discrimination, motivation and gender roles are taught, as are leadership theories (e.g., trait/behavioural, situational/behavioural, contingency). Research at the CMCs has also included studies on the integration of women (since 1980) at the CMCs. Gender integration issues are also addressed in varying ways at CF Staff School, Staff College, and National Defence College.

Conclusion

Gender integration has greatly expanded within the Canadian Forces, with respect to increased opportunities and actual employment of women in military occupations. Successful gender integration is a leadership responsibility. The CF is preparing its leaders at all levels to recognize and develop appropriate strategies to deal with many of the challenges of mixed-gender employment including physical differences, fraternization, prejudice and stereotypes, etc.

The recently initiated Combat Related Employment of Women (CREW) trials are taking this change process even further, by providing opportunities for training and employment of women in combat occupations and positions, on a trial basis. Lessons learned from SWINTER and the current mixed-gender experiences in combat service support units and ships are being incorporated into the preparation for and conduct of the CREW trials.

References

- Burns, J.M. (1978). Leadership. New York: Harper and Row.
- Department of National Defence (1978). The Officer. Winnipeg, Man.: Canadian Forces Training Materiel Production Centre.
- Lamerson, C.D. (1987). Integration of Women into Previously All Male Units: A Literature Review. (Working Paper 87-2). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.
- Personnel Research Associates (1988). Mixed Gender Service in Army Field Units: Leader's Considerations. Victoria, B.C.: Author.
- Personnel Research Associates (1988). Mixed Gender Service: Commanders' Considerations. Victoria, B.C.: Author.
- Personnel Research Associates (1988). Mixed Gender Service: Chiefs and Petty Officers. Victoria, B.C.: Author.
- Resch, Major G.D. (1988). Mixed Gender Leadership in Army Field Units. Briefing to Officers and N.C.O.'s of 1 Canadian Brigade Group, 12 January, 1988.

What Is This Thing Called Charisma?

Leanne E. Atwater, Ph.D.

United States Naval Academy

Robert Penn, Ph.D.

Linda Rucker, M.A.

San Diego State University

Many attempts have been made to distinguish true "leaders" from good managers. This leader/manager contrast has particular relevance in the context of military leadership. Janowitz (1959) articulated a concern that the military was shifting away from a focus on leadership to a focus on management. The leader was seen as being replaced, over time, by the military manager who would be responsible for complex military technology and who would draw on management strategies to get the business of the military done. Much discussion has ensued about whether leaders and managers are one in the same, or whether they are in fact two distinct sets of behavioral and personal characteristics. Segal (1981) contends that in recent years there has been an increasing tendency to use the terms management and leadership as synonymous. This development he finds disquieting. In his view, leadership refers to interpersonal processes while management is oriented toward maximizing profit. What is it that makes the difference between good management and good leadership? One important component of good leadership not usually considered characteristic of effective managers is charisma. Yukl (1981) describes charisma as one form of personal power. He asserts that while the exact qualities required for charismatic leadership are not well understood, a leader with charisma is better able to use inspirational appeals, and personal identification to influence subordinates. House (1977) contends that charismatic leaders excite, arouse and inspire their subordinates. Charismatic leaders radiate an impression of extreme confidence in their abilities as well as in their visions for the future.

Bass' discussion of transformational leadership, or leadership that goes beyond the transactional process, includes charisma as one of its critical components (Bass, 1987). He also sees charisma as a form of power and influence. Charismatic leaders inspire others' loyalty to the organization and have an ability to see what is important. They also inspire extra effort from subordinates—effort they are not likely to put forth for noncharismatic leaders.

Given the apparent value of charisma, it is a quality that the military as well as many other organizations seek in selecting and cultivating their leaders and future leaders. To achieve this goal, the components of charismatic influence must be ascertained. These components may include qualities of followers that increase their susceptibility to charismatic influence, or environmental factors conducive to the development and maintenance of charismatic leadership. But of seemingly primary importance are the qualities of leaders as perceived by subordinates that result in the leader being identified as charismatic. The present study focused on this fundamental component of charismatic leadership, namely, others' perceptions of the personal qualities of leaders that generate attributions of charisma. Also of importance in this study were the similarities and differences in perceptions of charisma in the military and civilian environments.

Historically, charisma has been a topic more pertinent to philosophers and sociologists than to empirical psychologists. Weber (1946) defined charisma as a quality of personality that conveys "extraordinariness" and the acquisition of "supernatural, superhuman or at least specifically exceptional powers or qualities."

Not until 1977 did an empirical model of charismatic leadership emerge. House (1977) extracted a model that identified components of charismatic leadership and their respective dynamics. This model also proposed methods to test the accuracy of its postulates. House's model is particularly significant because it appears to be the first major attempt to present charismatic theory in a form that lends itself to empirical examination. House's model transforms charisma or "extraordinariness" into three dimensions: leader characteristics, follower characteristics and situational characteristics. The postulate specifically relevant to the current study is the following: charismatic leaders can be described by a specific set of personal characteristics. The purpose of the present study was to determine the characteristics perceived by others that differentiate charismatic from noncharismatic leaders.

METHOD

Subjects were 85 volunteers (53 males and 32 females) whose ages ranged from 20 to 63 with a mean age of 32. The occupational breakdown was as follows: 35 military officers, 36 advanced university students, and 14 persons employed in various other capacities.

Each subject was asked to evaluate a charismatic and a noncharismatic leader of their choosing (no demographic or biographical constraints were imposed). Each chosen leader was then evaluated on 250 personally descriptive adjectives.

To ensure that responses reflected subjects' perceptions, the definition of charisma accompanying the questionnaire was purposely vague (i.e., "Charisma is a special, personal quality variously described as magnetism, charm, and/or extraordinariness that increases the individual's influence over others."

Questionnaire

Campbell's Work Orientation: Part IV, Descriptive Adjectives (Campbell, 1985) was used to obtain others' perceptions of the traits that distinguish charismatic from noncharismatic leaders. This scale was selected for the following reasons: its breadth as a descriptive tool, containing 250 adjectives; its capacity to measure degree of trait strength by using a Likert-type scale rather than a checklist approach to measurement; and the diminished ambiguity of trait definition it achieves by providing a brief description of each trait. The scale adjacent to each adjective contained six choices: Always, Usually, Sometimes, Occasionally, Rarely, and Never. An example of an adjective and its corresponding definition is "Busy - Occupied, never idle."

One adjective of particular importance was "Charismatic - has the ability to inspire others". This item was included to preserve the integrity of the scale as well as for use as a target adjective in correlational analyses with other adjectives.

Subjects were instructed first to "Name an individual who has occupied positions (often occupations) in which he/she projected charisma. This person should be someone you know through personal contact or through media exposure or readings (a public figure). The individual may be either alive or dead. However, it must be someone you know or know about sufficiently well to describe in some detail. If the individual is not well-known, please provide his/her occupation." Next, subjects were instructed to "Name an individual who has occupied positions (often occupations) that permitted the expression of charismatic qualities; however, this individual projected a lack of charisma. Use the same rules as above for selection." Finally the cover page, also, served to obtain certain biographical information about respondents, i.e. sex, age and occupation.

The scale containing 250 adjectives was administered twice to allow each subject to describe both a charismatic and noncharismatic leader. The instructions heading each form of the questionnaire were: "Write on the following line the name of the individual you identified as charismatic or noncharismatic. Then rate the frequency with which you feel that individual displays the following traits." The charismatic and noncharismatic instructions were counterbalanced with half of the subjects first describing the charismatic leader and half first describing the noncharismatic leader.

RESULTS

Frequencies tabulated on the sex and the occupations of charismatic and noncharismatic leaders nominated by subjects indicated that military officers primarily nominated male, military officers in their evaluations and civilians primarily nominated male, civilians. The occupation of civilian charismatic and noncharismatic leaders was primarily that of politician. Other civilian occupations of selected leaders included college professors, managers, and athletic coaches.

As a first attempt to determine which personal characteristics differentiated charismatic from noncharismatic leaders, matched pair t tests were performed on all 250 charismatic/noncharismatic pairs of adjectives. These analyses were not particularly enlightening in that 228 of the 250 adjectives were significantly different ($p < .05$).

A second approach involved correlating the target adjective (Charismatic - has the ability to inspire others) with the remaining adjectives. Items correlating with "Charismatic" .40 or higher are presented in Table 1. These items were considered the essential descriptors of charismatic leaders. As can be seen from Table 1, outgoing, sociable, cheerful and extroverted were most highly positively correlated with charismatic while shy pessimistic and gloomy were highly negatively correlated.

In all, 32 of the 249 trait adjectives were descriptive of charismatic leaders. Of these 32 items, 13 items clearly described the charismatic individual as sociable (e.g. outgoing, sociable, cheerful, extroverted), and 19 items unambiguously described the person as competent (e.g. resourceful, effective, resilient, bright, enterprising.)

Table 1
Analysis of Charismatic Leaders: Correlations with "Charismatic"

Correlation with Adjective "Charismatic"*		Correlation with Adjective "Charismatic"*		Correlation with Adjective "Charismatic"*	
Outgoing	.616	Skillful	.461	Enterprising	.406
Sociable	.550	Tender	.459	Likeable	.404
Cheerful	.530	Resilient	.453	Agile	.403
Extroverted	.523	Up-To-Date	.451	Agreeable	.401
Versatile	.506	Jovial	.442	Reliable	.401
Resourceful	.494	Well-Adjusted	.438	Shy	-.516
Effective	.490	Insightful	.435	Pessimistic	-.460
Inspiring	.486	Zestful	.432	Gloomy	-.452
Charming	.483	Friendly	.428	Aloof	-.406
A Leader	.480	Secure	.419	Quiet	-.406
Dynamic	.478	Bright	.410		

* All correlations are significant ($p < .001$)

As can be seen from Table 2, noncharismatic individuals were seen as aloof, private, introverted, Not charming, Not creative, Not comic, and Not inventive, among other undesirable characteristics. Interestingly, about half of the characteristics that describe charismatic individuals also describe noncharismatic individuals in their absence. For example, charismatic leaders are charming while noncharismatic individuals are Not charming. Traits, the absence of which describe only noncharismatic individuals and the presence of which are not descriptive of charismatic individuals include creative, comic, inventive, witty, far-sighted, private and introverted.

Table 2
Analysis of Noncharismatic Leaders:
Items Most Highly Correlated with the "Charismatic" Item

Adjective	Correlation with "Noncharismatic" *
(Not) Charming +	.559
(Not) Creative	.495
(Not) Comic	.491
(Not) Inventive	.461
(Not) Inspiring +	.454
(Not) Versatile +	.451
(Not) Witty	.443
(Not) Zestful +	.442
(Not) Friendly +	.434
(Not) Far-Sighted	.429
(Not) Curious	.412
(Not) Outgoing +	.401
Aloof	-.466
Private	-.418
Introverted	-.403

* Statistically significant at $p < .001$ level.

+ The presence of this trait describes charismatic leaders.

In summary, the analyses presented thus far revealed three significant patterns. First, charisma is described primarily by adjectives indicative of sociability and competence. Second, noncharismatic individuals are described primarily by an absence of positive characteristics rather than a preponderance of negative attributes. (A number of negative adjectives were among the 250 included in the scale. These were not often descriptive of noncharismatic leaders). Third, noncharismatic leaders are clearly viewed as lacking social skills.

A second purpose of this study was to investigate military and civilian perceptions of charisma. Is charismatic leadership characterized by the same attributes in the eyes of military officers as it is by civilians? To address this question, responses by military and civilian subjects were compared. Because military officers chose primarily other officers as their target leaders this analysis should provide insight as to just what charisma means in the military community, both in terms of raters and their chosen leaders. T-tests were performed between military and civilian responses on each descriptive adjective. Thirty-two of the items differed significantly ($p < .05$). These items are presented in Table 3.

Table 3
Analysis of Charismatic Leaders: Means, t-test Values and Significance
Levels Between Military and Civilian Subjects' Evaluations

Adjective	Military Mean	Civilian Mean	t-Value	p-Level
Demanding	4.86	3.73	3.98	.000
Obedient	4.65	3.65	3.66	.000
Rugged	5.00	4.00	3.50	.001
Traditional	4.15	3.36	3.32	.001
Handy	4.26	3.14	3.26	.002
Prudent	4.91	4.21	2.91	.005
Meek	1.68	2.33	-2.85	.006
Mechanical	4.28	3.35	2.79	.007
Inventive	5.00	4.53	2.71	.008
Brave	5.31	4.83	2.68	.009
Conforming	3.63	2.94	2.68	.009
Domestic	4.69	3.81	2.62	.011
Conventional	4.37	3.69	2.57	.012
Far-Sighted	4.77	5.29	2.47	.016
Rebellious	2.25	3.00	-2.46	.016
Mathematical	4.77	4.07	2.43	.018
Dependable	5.46	5.04	2.39	.019
Persuasive	5.53	5.16	2.32	.023
Responsible	5.65	5.33	2.29	.025
Fickle	1.88	2.35	-2.28	.025
Strict	4.47	3.92	2.27	.026
Unflappable	5.12	4.63	2.25	.027
Well-Educated	5.62	5.22	2.22	.029
Sensible	5.44	5.13	2.21	.030
Comic	4.05	4.63	-2.19	.032
Gentle	4.03	4.65	-2.17	.033
Risk-Taking	4.68	4.15	2.14	.036
Punctual	5.29	4.83	2.12	.037
Frugal	4.36	3.76	2.11	.038
Purposeful	5.50	5.23	2.10	.039
Ethical	5.38	4.94	2.08	.041
Wholesome	5.29	4.79	2.07	.041

Of particular interest are the directions of the differences. Military officers rated charismatic leaders higher than did the civilian raters on 14 of the items including the following: tough (rugged, brave), responsible (dependable, punctual), conservative (prudent, traditional, obedient, conforming), ethical (wholesome), innovative (inventive, far-sighted), and determined (demanding, risk-taking, purposeful). For example, the average level of "demanding" was 4.86 for the military sample and 3.73 for the civilian sample. A number of these descriptive differences are consistent with findings that those naval officers seen as very high on masculinity were more likely to receive recommendations for early promotion in their careers than their non-masculine counterparts (Berry and Baker, 1989). The only adjectives which were less descriptive of military charismatic leaders than of civilian charismatic leaders were meek, far-sighted, rebellious, fickle, comic and gentle. These were not primary components of charisma in either group.

For the most part, military officers define charismatic leadership very similarly to their civilian counterparts, with a number of added dimensions. When analyzed as separate samples, both civilian and military raters saw charismatic individuals as possessing very different traits than noncharismatic individuals. (The analyses of the two samples separately paralleled results for the total sample). Consequently, not only do both the military and civilian definitions of charisma include sociability and competence, charisma in the military includes two additional dimensions. These dimensions can be described in general terms as a masculine/rugged dimension and a conservative/responsible dimension. The masculine traits include demanding, mechanical, brave, not gentle and risk-taking. The conservative descriptors include traditional, prudent and wholesome. (See Table 4.)

Table 4
Categorization of Additional Traits Descriptive of Charisma
in a Military Environment

<u>Rugged/Masculine</u>	<u>Conservative/Responsible</u>
Demanding	Traditional
Rugged	Obedient
Handy	Prudent
Not Meek	Conforming
Mechanical	Conventional
Brave	Dependable
Mathematical	Responsible
Strict	Sensible
Unflappable	Punctual
Not Gentle	Frugal
Risk Taking	Purposeful
	Ethical

In conclusion, the results suggest that individuals have a fairly clear and consistent idea of traits indicative of charisma. Many of these traits are common across civilian and military communities. In the military, however, additional traits are required for an individual to be seen as clearly charismatic. These additional traits seem particularly relevant to war-fighting capabilities and to the ethical standards needed to best defend and represent our country.

DISCUSSION

The results from this study suggest that a number of common characteristics are descriptive of charismatic leaders in different sectors. These characteristics include being sociable, competent, friendly and well-adjusted. Many of the qualities descriptive of charisma appear to be more amenable to selection than to training if one is interested in charismatic leadership qualities. Others, such as being skillful and up-to-date are clearly training issues. Bass (1987) and his colleagues report success in training leaders to be more charismatic.

Of special interest in the military context are the qualities demanded of military leaders considered charismatic. While rugged qualities depicting the classic warrior are important in charismatic military leadership, the ethical component is also important. Campbell (1987) in his efforts to define leadership traits (or leaders' orientations to the world as he puts it) has been unable to identify a dimension depicting integrity or ethics. Charisma in his model is included in a broader category he calls "dynamic". It may be that ethics or integrity are important only in some leadership contexts, or perhaps integrity is of less particular importance to leadership in general and emerges as important to the charismatic component of leadership. Regardless, it certainly arises in military officers' perceptions of what is charismatic leadership.

In sum, this study has addressed one of the postulates in House's model (House, 1977), identifying the personal characteristics descriptive of charismatic leaders. It has also suggested that charisma may have additional descriptors in particular environments, namely the military. Secondarily, it has illuminated questions concerning the relevance of qualities such as a sense of ethics, responsibility and conformity to effective military leadership. The question arises as to the durability of the descriptors of charisma over time. Perhaps what is seen as charismatic differs significantly as the eyes of the beholders' values adapt to a changing society. A follow-up study in 10 years would be interesting. Also valuable would be an extension of this study into the worlds of politicians, business managers, entertainers, and religious leaders, as well as an amplification in the military and professional arenas. Additionally, analyses of charismatic qualities as a function of the sex and age of the raters would prove insightful. Future research is most definitely needed to provide a more thorough understanding of this complex human quality.

References

- Bass, B.M. (1987). Policy implications of a new paradigm of leadership. Paper presented at the 1987 Navy Leadership Conference, Annapolis, MD.
- Berry, V. and Baker, H. (1988). The relationships of physical attractiveness and perceived masculinity to the performance evaluation and occupational success of naval officers. Paper presented at the Eleventh Biennial Psychology in the Department of Defense Symposium, April, 1988.
- Campbell, D. (1985). Campbell Work Orientations: Part IV, Descriptive Adjectives. Greensboro, N.C.: Center for Creative Leadership.
- Campbell, D. (1987). Leadership performance. Paper presented at the 1987 Navy Leadership Conference, Annapolis, MD.
- House, R.J. (1977). A 1976 theory of charismatic leadership. In J.G. Hunt & L.L. Larson (Eds.), Leadership: The cutting edge. Carbondale, IL: Southern Illinois University Press.
- Janowitz, M. (1959). Changing patterns of organizational authority. Administrative Science Quarterly, 3, 474-493.
- Segal, D. (1981). Leadership and management: Organization theory. In J.H. Buck and L.J. Korb (Eds.), Military leadership. Beverly Hills, CA: Sage Publications.
- Weber, M. (1946). The sociology of charismatic authority. In H. Gerth & C.W. Mills (Eds.), From Max Weber: Essays in Sociology. New York: Oxford University Press.
- Yukl, G.A. (1981). Leadership in organizations. Englewood Cliffs, N.J.: Prentice-Hall, Inc.

THE EFFECTS OF INDIVIDUAL EXPERIENCE AND INTELLIGENCE ON DECISION PERFORMANCE

Jody C. Locklear, Charles G. Powell and Fred E. Fiedler

UNIVERSITY OF WASHINGTON

ABSTRACT

A field experiment in a military setting assessed the contribution of intelligence and experience on pre-decision information search processes under three conditions of stress. The comparatively more intelligent leaders searched for more information and repeated their inspection of information items more often, and they out-performed less intelligent leaders in all conditions. Experienced leaders searched for less information but performed better in the high stress condition. The results are discussed in the context of Cognitive Resource Theory and their implications for military planning and decision-making in stressful environments.

INTRODUCTION

Research on decision-making has focused on highly controlled experiments rather than the complex organizational settings in which decisions usually are made. Partly for this reason, decision research has tended to ignore individual differences and their effect on the search for relevant information as well as the communication and implementation of plans and decisions.

Recent research (Payne, 1982 and Nichols-Hoppe, 1984) indicates that decision processes are highly contingent upon characteristics of the decision task, and Hogarth and Einhorn (1981) have remarked that "The most important empirical results have shown the sensitivity of judgement and choice to seemingly minor changes in tasks." In addition, Beach and Mitchell, 1978; and McAllister, Beach & Mitchell, 1979, have provided evidence that such other factors as the situation's characteristics also will influence decision processes. Beyond these studies, however, there has been little research on the effects of such moderating variables as stress, intelligence, and experience on the decision process which are the focus of the present study.

Individual Characteristics

Payne (1976) observed in an informal aside that the manner in which information is processed during decision tasks seems to vary considerably among individuals. This point is supported by Weiss and Knight (1980) who investigated the relationships among self-esteem, information search, and problem-solving efficiency. They reported that individuals with low self-esteem (a) devoted more effort to searching for information, made their search more functional, and (c) were more effective in the task than persons high in self-esteem. The subjects in the Weiss and Knight study, however, did not deal specifically with decision processes but rather to determine a rule by which three numbers were related.

That such obvious personal characteristics of decision-makers as intellectual abilities and previous job-related experience are important has already been shown in research by Fiedler and his associates (see Blades, 1967; Borden, 1980;

Fiedler, Potter, Zais & Knowlton; Potter & Fiedler, 1981). These investigators found that the leader's intellectual abilities contribute to the task only when the leader reports low stress. The research suggests, therefore, that the performance of groups and organizations can be substantially improved simply by more effectively using the abilities and knowledge which already are available in groups and organizations (e.g., Blades, 1986; Fiedler, Fiedler & Garcia, 1987; Fiedler, Potter, Zais & Knowlton, 1979).

Cognitive Resource Theory

Our research has led to the development of a model, called Cognitive Resource Theory, that is based on Blades' research (1976) and a number of diverse studies subsequently conducted on military personnel (e.g., Borden, 1980; Fiedler, Potter, Zais & Knowlton, 1979; Potter & Fiedler, 1981). Cognitive Resource Theory is designed to explicate the conditions under which leaders' cognitive abilities, and the resulting concrete behaviors, promote or block successful group performance.

One disruptive variable of major importance is the stress the leader experiences in the relationship with his or her superior--"boss stress" (Borden, 1980). In somewhat oversimplified form, when boss stress is high, leaders rely on their experience by not their intelligence; when it is low, they rely on their intelligence but not their experience (e.g., Fiedler, Potter, Zais & Knowlton, 1979; Potter & Fiedler, 1981).

Boss stress, in particular, seems to affect the leader's ability to use intelligence in making sound decisions (Fiedler, Potter & McGuire, 1988). We view the decision-process as consisting of the recognition of the problem, a search for relevant information and processing of the information, formulation of the decision, and implementation of the decision. We now ask how one particular aspect of the decision-process, namely the search for relevant information, is affected by stress and the leader's intellectual abilities and experience. Specifically, do leaders of (a) relatively high or low intelligence, (b) high or low levels of experience, and (c) working under conditions of relatively low stress, job or interpersonal stress, utilize different information-search strategies in making their decisions.

Our current working model of the decision process is that stress inhibits the effective utilization of intelligence preventing the individual from thinking of the appropriate response, or from narrowing down the options. Experience is seen as a filter that reduces the possible alternatives to a workable number and thereby enabling the individual with even comparatively modest intellectual skills to select a viable alternative.

METHOD

Subjects. Lieutenants (n=25), captains (n=12) and platoon sergeants and above (n=25) from a combat arms battalion of an active Army division participated in this study.

Design. The basic design of the experiment included three independent variables. Two of these variables involved between-subject factors, and these were completely crossed, forming a 2 X 2 factorial design. Two levels of intelligence and experience were determined by median splits. Every subject performed the task under three experimentally induced conditions of stress.

Task and Procedure. The task consisted of a modified version of an army training activity, namely, the preparation of a weekly training schedule. Subjects were informed that they were participating in the development of a new Training Manual (TM) describing how to prepare schedules for the new modernized Army. After the initial briefing they were required to complete intelligence, job knowledge, and biographical data sheets as a part of the Army's validation data. Time in service was the measure of experience.

After completing the scales, subjects were randomly assigned to 3-man groups and dismissed. Upon returning later that day, each 3-man group was shown the demonstration board while the experimenter explained briefly that each person could (a) read as much or little of the information presented on the board as he wished; b) look at only one piece of information at a time and replace the information card before turning to the next; c) return to the same item of information as many times as he wished; but d) that the time allowed to rank the schedules would be indicated by the monitor for that condition.

Each subject was then randomly assigned to a condition and rotated through the remaining two conditions in counterbalanced order. In the "base-rate" condition, the subject worked under relatively little pressure, and was reassured that he had sufficient time for the task. In the "time constraint" condition, stress was induced by informing the subject that he had only half the time required by the average subject to complete the task. In the "evaluative apprehension" condition, stress was induced by having the subjects perform the task in front of a video camera, and telling them that they might be observed by their battalion commander or sergeant's major. Upon completion of the tasks under the three conditions and the post-test, Ss were debriefed and returned to their companies.

Monitoring information acquisition. The dependent variables consisted of observations of the individual's behavior in using an "information board" (Wilkins, 1967; Payne, 1976) for making the three decisions. Each information board contained 35 3x5 items of information mounted on cards about 5 possible training schedules. The schedules were the rows on the information board. The 7 categories/attributes for each schedule were considered critical for the task by a panel of experts, and these categories formed the columns of the information boards. Subjects were free to consult the information board for any amount of information and as often as they liked prior to rendering their decision. They were informed that the information on the board was accurate. A trained recorder noted which items each subject selected prior to making the decision.

The information board yielded four dependent measures: 1) number of information items were inspected during the process of information search, 2) how often the same items was reinspected, 3) to what degree the ranking was similar to that of a panel of experts and 4) what types of items were preferred across all conditions.

Stress Manipulation. Three stress conditions were experimentally induced. In the "low stress" condition subjects were assured that they had plenty of time and that the task was not critical. In the "time constraint" condition, subjects were told that they would have only half the time required on the average to complete the problem. In the "evaluation apprehension stress" condition subjects performed the task in front of a TV camera and they were told that their performance might later be evaluated later by a panel of experts which included their battalion sergeant major and commander.

RESULTS

Manipulation Checks. A State Anxiety Inventory (SAI) was administered during the initial briefing prior to any other testing or exposure to the decision tasks, and again on completion of the stress condition. A 2-way Analysis of Variance (ANOVA) revealed that subjects in the "evaluative apprehension stress" condition perceived higher levels of stress ($F(1,56)=5.47, p < .001$).

Amount of Information Inspected. A multi-variate analysis of variance compared the number of information items that subjects inspected in each of the three conditions. This analysis showed first, that more intelligent leaders viewed more items of information in all conditions than did less intelligent leaders ($F(2,46)=8.49, p < 0.01$).

Second, the more experienced leaders examined considerably fewer information items than did inexperienced leaders ($F(2,46)=14.75, p < .001$). These results suggest, that in general the more intelligent and less experienced leaders wanted to make sure of their decisions by double checking information and including all available information in their decision process. The more experienced individuals acted as if they did not need as much information to make their decision.

Task performance. A multi-variate analysis of the performance scores disclosed that (a) the more intelligent subjects performed better in all 3 stress conditions ($F(2,46) = 8.74, p < 0.01$); (b) experience interacted with the experimental condition: As expected on the basis of previous studies, the more experienced leaders out-performed the less experienced leaders in the evaluative stress condition (though not under time stress). Third, the more experienced subjects viewed significantly fewer items than did their less experienced counterparts. However, in the stress condition the more experienced leaders seemingly chose the "right" information, i.e., the information considered critical by experts. In the two less stressful conditions, experienced and inexperienced subjects did not differ significantly in their performance score.

Pattern of Information Acquisition. Search Organized around specific Attributes. The only main effect observed involved experience: the more experienced individuals, in general, selected different information in their pre-decisional search than less experienced

($F(2,96) = 9.747$, $p < 0.001$). Specifically, the inexperienced individuals searched more often for the boss' guidance/approval and what was termed "peace time" mission requirements, while the experienced leaders inspected information on how well the training schedule incorporated unit evaluation guidelines and "war time" mission requirements.

DISCUSSION

The results of this study showed that the more intelligent leaders chose more information items and performed better on the ranking task than did relatively less intelligent leaders. As mentioned earlier, we view the decision process as consisting of problem recognition, information search and processing, and formulation of the decision or plan. If this analysis is correct, it seems that neither time constraint nor evaluation apprehension stress affected the information search phase of the decision process.

We must also note, however, that the ranking task was performed more effectively by the more intelligent than less intelligent leaders, regardless of stress. Previous research had led us to believe that more intelligent leaders would perform better in the low stress and the job stress conditions, but less well in the evaluation-stress condition. While the results were in the expected direction, they fell far short of significance, either because the effects were weak or because sample was quite small.

We had expected the more experienced leaders to perform better than less experienced leaders in the stress condition. This was the case only in the evaluation stress conditions but not in the time constraint condition. Again, since none of the differences were significant, they have to be considered at best as suggestive. The results do allow us to conjecture, however, that the more experienced leaders knew what to look for, and did not have to deal with more information than was essential for making their decision. We now need to consider further what kinds of choices the more intelligent leader makes under stress that result in poorer performance, and the choices of the more experienced leader under the time constraint and low or base-rate condition that result in poorer decisions than one would normally expect to find under these conditions.

REFERENCES

- Beach, L.R., & Mitchell, T.R. (1978). A contingency model for the selection of decision strategies. *Academy of Management Review*, 3, 439-449.
- Blades, J. W. (1976). The influence of intelligence, task ability, and motivation on group performance. Unpublished doctoral dissertation, University of Washington, Seattle.
- Bordon, D. F. (1980). Leader-boss stress, personality, job satisfaction and performance: Another look at the interrelationship of some old constructs in the modern large bureaucracy. Unpublished doctoral dissertation, University of Washington, Seattle.

Einhorn, H.J., & Hogarth, R.M. (1981). Behavioral decision theory: Processes of judgment and choice. Annual Review of Psychology, 32, 53-88.

Fiedler, F. E., Potter, E. H., Zais, M. M., & Knowlton, W. A., Jr. (1979). Organizational stress and the use and misuse of managerial intelligence and experience. Journal of Applied Psychology, 64, 635-647.

Fiedler, F. E. & Garcia J. E. (1987). New Approaches to Effective Leadership: Cognitive Resources and Organizational Performance. New York: Wiley.

Hunt, E. (1985). Verbal Ability, in Human Abilities: An Information Processing Approach. 31-58, San Francisco: Freeman.

McAllister, D.W., Mitchell, T.R., & Beach, L.R. (1979). The contingency model for the selection of decision strategies: An empirical test of the effects of significance, accountability, and reversibility. Organizational Behavior and Human Performance, 24, 228-244.

Nichols-Hoppe, K. T. (1984). Information Acquisition as a Contingent Decision Process. Unpublished doctoral dissertation, University of Washington, Seattle.

Payne, J.W. (1976). Task complexity and contingent processing in decision making: An information search and protocol analysis. Organizational Behavior and Human Performance, 16, 366-387.

Payne, J.W. (1982). Contingent decision behavior. Psychological Bulletin, 92(2), 382-402.

Potter, E. H., & Fiedler, F. E. (1981). The utilization of staff member intelligence and experience under high and low stress. Academy of Management Journal, 24(2), 361-376.

Spielberger, C.D., Gorsuch, R.L., & Lushene, R. (1968). Self-evaluation questionnaire. Palo Alto, CA: Consulting Psychologists Press.

Weiss, H.M., & Knight, P.A. (1980). The utility of humility: Self-esteem, information search, and problem solving efficiency. Organizational Behavior and Human Performance, 25, 216-223.

Wilkins, L. (1967). Social deviance. Englewood Cliffs, NJ: Prentice Hall.

The Effect of Stress on the Performance of Creative and Intelligent Leaders

Fred E. Fiedler and Frederick W. Gibson

University of Washington

Our research over the past 17 years has investigated the conditions under which the intellectual abilities of leaders and members are effectively used in the performance of organizational tasks (Blades, 1976; Fiedler & Garcia, 1987). This work has led to a "cognitive resource theory" (Fiedler and Garcia, 1987) that attempts to systematize these and related findings on the role of cognitive resources in leadership. This research shows, for example, that the leader's intelligence or creativity contribute most strongly to group performance if the leader is directive and free from stress, or is directive and has group support. The theory proposes that the leader must concentrate on the task in order to make good plans and decisions, and that these plans/decisions must subsequently be communicated to a motivated group which is willing to implement them.

As Sarason (1984) and others have shown, stress appears to distract the leader by introducing irrelevant thoughts that interfere with task performance. The leader finds it difficult or impossible to concentrate on the intellectual aspects of the task. As a result, we would expect the correlation between leader cognitive resources and group performance would be near zero under these conditions.

However, somewhat unexpectedly, we find in a number of these situations that leaders' intellectual abilities are negatively correlated with performance. These negative correlations are too strong and too frequent to be brushed aside as chance findings. For example, the correlation between leader intelligence and supervisor's ratings of performance for platoon leaders was .64 ($n=22$; $p<.01$) when stress was low, but was $-.20$ ($n=30$) when stress was high (Borden, 1981). Similarly, in a study of 130 Coast Guard officers and petty officers, correlations between intelligence and job performance ratings for the most intellectually demanding tasks under conditions of low boss stress was .27 but was $-.46$ when reported boss stress was high (Potter and Fiedler, 1981).

It is easy to understand why more intelligent leaders might perform no better than less intelligent leaders when stress is high. It is more difficult to explain why brighter or more creative leaders perform less well than less bright or creative leaders when distracted by stress. One plausible explanation is that more intellectually gifted people have higher expectations of their ability to cope with intellectually demanding tasks and are more stimulated by these tasks and therefore motivated to do well. For this reason, individuals with high ability are likely to (a) take a more active part in attempting to deal with these

tasks; that is, they will talk more; (b) strain for more exotic, original, or elegant solutions (and withhold those ideas that don't initially measure up to these standards, even though they might well work); and (c) by talking more reduce the opportunity of other group members to contribute to task accomplishment.

These hypotheses find some support in an earlier study of group creativity (Fiedler, Meuwese & Oonk, 1961). This study revealed a correlation between number of comments on the part of group members and performance of $-.72$ ($n=32$) in groups which were relatively tense and stressful, as indicated by the presence of a person who was seen as a destructive critic. It appears, therefore, that the sheer amount of talking detracts from performance. Furthermore, there was a substantial yet nonsignificant correlation between leader intelligence and amount of talking in the more stressful condition ($r=.70$; $n=10$) but not in the nonstressful condition ($r=.02$; $n=10$).

This discussion suggests that more creative or intelligent leaders (a) talk more under stress than their less gifted counterparts; (b) generate fewer ideas (due to higher standards they create for each one; i.e., they try to "hit home runs"); and (c) reduce the chance for contribution by group members; here, operationalized as less talking and introduction of fewer ideas on the part of the members.

METHOD

Data for this study were obtained in a laboratory experiment conducted by Meuwese and Fiedler (unpublished technical report, cited in Fiedler & Garcia, 1987, p 149-151). This experiment investigated the effect of stress on the behavior and performance of task-motivated and relationship-motivated leaders. Fifty-four three-man groups of Reserve Officer Training Corps (ROTC) cadets participated in two tasks: (a) proposing a new pay plan for the various ROTC programs (which at the time gave different stipends to the cadets of the three services) and (b) inventing a fable for school children. Post-session questionnaires measured participants' state anxiety (Alexander and Husek, 1962). The Multi-apitude intelligence test (Psychological Corporation) and the Guilford-Christenson Plot Titles and Alternative Uses tests of creativity (Guilford, Berger, and Christenson, 1954) assessed subjects' intelligence and creativity, respectively. The group products were rated by three judges in the case of the pay proposal and five in the case of the fable.

One analysis of the data determined the overall contribution of leader cognitive resources to group performance. The groups were divided into those in which the leader rated the situation as low, moderate, or high in stress, and each leader's creativity and intelligence scores were then correlated with performance. As in previous studies, we found positive correlations between both leader creativity and intelligence with group performance for leaders perceiving low stress ($r .31$ and $.42$, respectively).

Under high felt stress, we found negative correlations between leader creativity and performance and zero correlation between leader intelligence and performance ($r = -.23$ and $.05$, respectively). Stress therefore did not substantially moderate the relationship between leader intelligence and group performance but did so for the leader creativity - group performance linkage. We therefore focus on this relationship.

Typed transcripts of each session were content analyzed by three independent judges. This analysis was preceded by a one-hour rater error training session for the judges in which definitions and operationalizations of "talking" (the number of lines in the transcript associated with comments by the leader or the rest of the group) and "ideas" (any original substantive comment by the leader or members which aided in the attainment of the group goals or solving the task-related problem) were presented and discussed.

Raters labelled and counted four items: the number of lines of talk by the leader; the number of lines of talk by the group (excluding the leader); the number of ideas presented by the leaders; and the number of ideas presented by the group. Thus we obtained four measures for each group task. Following completion of ratings of three transcripts, a followup rater training session was held in which disagreements and misunderstandings were discussed and resolved to the satisfaction of all members.

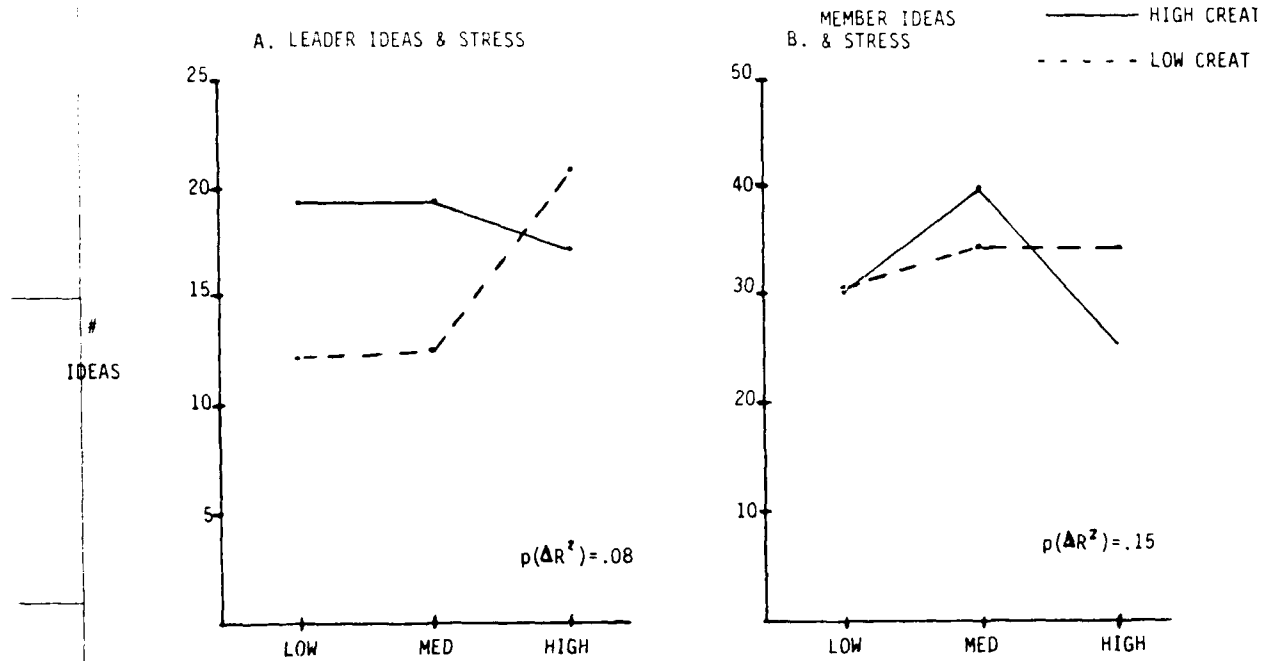
RESULTS

Inter-rater agreement was assessed by computing the Pearson correlation between all pairings of ratings of the four variables described above. Overall inter-rater agreement across the four variables was $.87$. Test-retest reliability after an average time lag of five weeks was $.97$. Both indices indicate substantial stability of the ratings across the two dimensions.

To determine the effects of stress on the variables under study, several moderated regression analyses were conducted in which the dependent variable was one of the four rated indices and the independent variables were leader creativity (or intelligence), rated stress, and the product of these two terms (representing the moderator term). The incremental change in R^2 due to the interaction was then tested for significance. To visualize the nature of the ability-by stress interactions, cell means corresponding to three levels of felt stress and two levels of leader creativity or intelligence were plotted. Figure 1 contains the plotted cell means for the moderated regression analyses.

The major findings are that leader creativity effected only the number of ideas generated by the leader or group members. On the other hand, leader intelligence effected only the amount of talking. Consequently, we discuss only these effects.

LEADER CREATIVITY RESULTS



LEADER INTELLIGENCE RESULTS

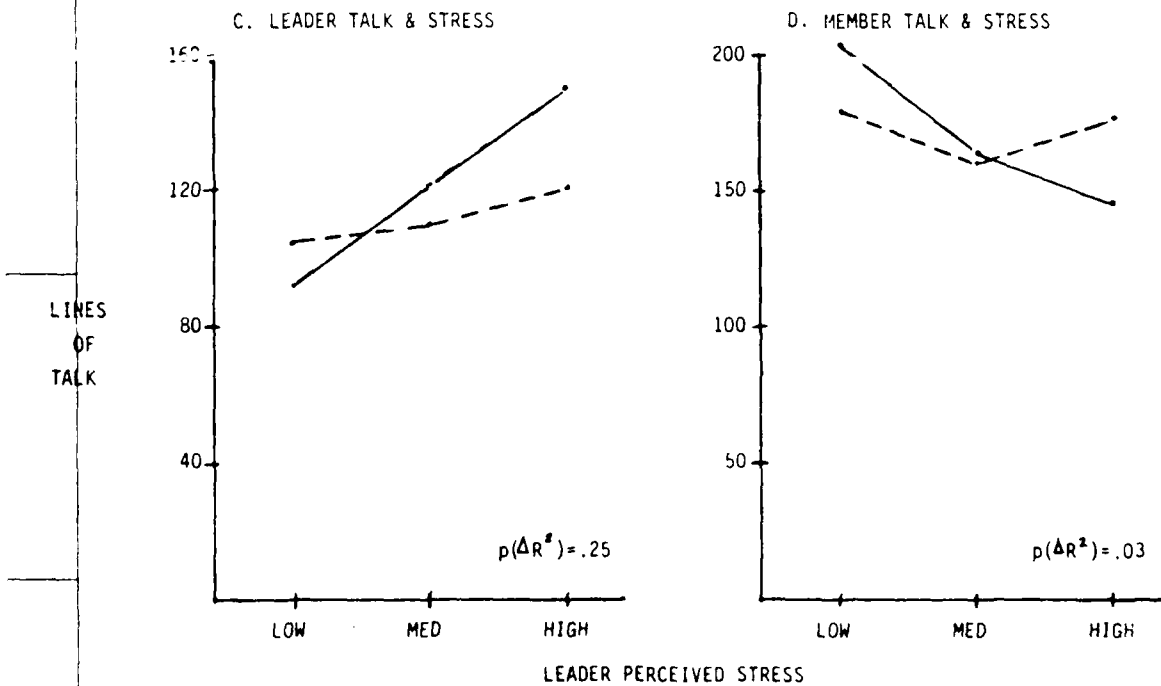


FIGURE 1: SUMMARY OF MAJOR FINDINGS

Creativity. As Figure 1A indicates, under low perceived stress, creative leaders offered more ideas than did less creative leaders. However, this trend was reversed for leaders who felt high stress. The p value for the interaction term describing this relationship ($p(\Delta R^2)$) was .08. Figure 1B displays the number of ideas offered by group members whose leaders were less or more creative. Under low stress, there was no difference in the amount of ideas generated. Under high stress, though, groups with creative leaders offered fewer ideas than did groups of less creative leaders. This interaction is not as strong, but it is marginally significant ($p(\Delta R^2)=.15$).

Intelligence. Figures 1C and D indicate that under low stress, intelligent leaders talked less than less intelligent leaders, whereas their members talked more; under high stress these trends reversed. This interaction is marginally significant for leaders ($p(\Delta R^2)=.25$) but stronger for group members ($p(\Delta R^2)=.03$).

DISCUSSION

We may summarize our findings as follows: Stress and leader creativity interact to suppress the generation of task-related ideas on the part of both the creative leader and his group members. On the other hand, stress and leader intelligence interact to "cause" intelligent leaders to talk more under stress, but their group members to talk less. Explanations for these findings are highly speculative, but we suggest that for creative leaders, the motivation to "hit home runs" under stress creates such expectations for each idea that very few are deemed worthy of mention. Somehow, these expectations are communicated to group members, who subsequently offer fewer ideas, too. For intelligent leaders, stress is associated with more talking, perhaps because of felt pressure to excel, whether internally generated or communicated by the group. Regardless of source, this increased talkativeness suppresses discussion by the rest of the group. Since both idea generation and evaluative discussion seem necessary for effective performance on intellectually demanding tasks, the behavioral tendencies associated with creative and intelligent leaders under stress imply deleterious consequences for group performance.

Cognitive resource theory proposes that leaders must concentrate on the task to make good plans/decisions, and that these plans/decisions must be communicated to a motivated group willing to implement them. This study implies that under stress, gifted leaders and perhaps their members behave in such a way as to break this "performance chain": leader creativity seems to effect the planning stage, whereas leader intelligence may effect the elaboration of ideas.

These are important findings when we consider that leaders are often selected and assigned in part on the basis of their intellectual abilities. Moreover, those who appear most able are

typically asked for advice and assistance in making major decisions. If these individuals are considerably less effective under certain conditions than those who are intellectually less able, it is clear that the organization will experience major difficulties. To the extent that we can identify the conditions and mechanisms under which these more able leaders will be dysfunctional, however, we can design more effective selection and assignment policies. Should these programs be inflexible, we can at least educate our leaders in the basics of group process, identify their behavioral tendencies to them, and train them to recognize and resist the effects we have identified here.

References

- Alexander, S., & Husek, T.R. (1962). The anxiety differential: Initial steps in the development of measures of situational anxiety. *Educational and Psychological Measurement*, 22, 325-348.
- Blades, J.W. (1976). The influence of intelligence, task ability, and motivation on group performance. Unpublished doctoral dissertation, University of Washington, Seattle.
- Borden, D.F. Leader-boss stress, personality, job satisfaction and performance: Another look at the inter-relationship of some old constructs in the modern large bureaucracy. Unpublished doctoral dissertation, University of Washington, Seattle.
- Fiedler, F.E., & Garcia, J.E. (1987). New Approaches to Leadership: Cognitive Resources and Organizational Performance. New York: John Wiley.
- Fiedler, F.E., Meuwese, W.A.T., & Oonk, S. (1961). An exploratory study of group creativity in laboratory tasks. *Acta Psychologica*, 18, 100-119.
- Guilford, J.P., Berger, R.M., & Christenson, A. (1954). A factor analytic study of planning: Vol 1: Hypothesis and description of tests. University of Southern California, Psychological Laboratory, Los Angeles, CA.
- Potter, E.H., & Fiedler, F.E. (1991). The utilization of staff member intelligence and experience under high and low stress. *Academy of Management Journal*, 24(2), 361-376.
- Sarason, I.G. (1984). Stress, anxiety, and cognitive interference: Reactions to tests. *Journal of Pers and Social Psych*, 46, 929-938.

COST-EFFECTIVE RETENTION OF ARMY WEST POINT OFFICERS ¹

Hyder Lakhani, U.S. Army Research Institute, Alex., VA.

Rashmi Lal, Engineering and Economic Research Inc., Reston, VA

Retention of U.S. Military Academy (USMA) or West Point trained junior officers has been a subject of considerable interest to the U.S. Army management (Lakhani, 1988). There was also a debate in the Armed Forces Journal International and the Army Times whether the USMA should be abolished because it was argued that the USMA is not cost-effective in training and retaining high quality officers. Therefore, an objective of this paper is to analyze cost-effective retention of USMA and Reserve Officer Training Corp (ROTC) junior officers. The next Section reviews the literature and data on problems of retention of these officers. Section 3 describes our simultaneous equations model of retention and the database for an empirical model. Section 4 presents the results and concludes that retention of USMA officers is cost-effective for the Army.

REVIEW OF LITERATURE AND DATA

Schemmer (1985) proposed abolition of the USMA because of its high cost of \$225,000 per Second Lieutenant compared to only \$5,000 per ROTC officer and a declining "trend" in quality of USMA officers. The soldier quality was measured by Schemmer in terms of the percentage of USMA officers in the general officers strength in the years 1984 and 1985, e.g. of the 64 Brigadier Generals selected last year, only 23% were West Point graduates. Scott (1985) rebutted Schemmer by noting that the cost of USMA education, according to the U.S. General Accounting Office, was about \$175,000 in 1984 and that the USMA training was superior to ROTC because their academic load was 152 credit hours instead of 120 to 140 for ROTC. Doleman (1985) also rebutted Schemmer by saying that the data on general officers for only two years was not a "trend". Doleman added that the quality of USMA applicants was superior because 83 percent of them were in the top of their classes. Schemmer (1985) responded to Doleman by presenting 1965-85 data on general officers and noted that the percent of USMA graduates had a declining trend. A problem with Schemmer's trend data is that they are not adjusted for the increase in non-USMA officers over the twenty years. For example, the number of new non-USMA officers have been about eight times (8,000 versus 1,000 per year) that of new USMA officers (Scott, 1985). Therefore, since 11% USMA graduates made up 23% of Brigadier Generals the quality is not inferior. Adelsberger (1988) reviewed the cost controversy and referred to a study by Fagan who adjusted Schemmer's lower cost of ROTC officers by

¹ The views expressed in this paper are solely those of the authors and do not reflect the views of the Dept. of the Army or the U.S. Army Research Institute.

adding the subsidies paid by the States to ROTC colleges and concluded that ROTC and USMA costs were comparable at \$183,000 and \$191,000 respectively.

The literature on retention reveals that Gotz and McCall (1983) analyzed the stay/leave decisions of aggregation of Air Force officers and concluded that retentions were significantly related to promotion opportunities, retirement benefits and civilian and military income opportunities. Smith (1987) concluded that an Air Force officer whose spouse was in the labor force was less likely to stay. Johnston (1988) concluded that the separate models by commission group predicted retention better than an aggregate model of all junior officers. These empirical results are used as the theoretical foundations in our interdisciplinary model.

The data on attrition of officers show that attrition of junior officers is a problem because 30 percent of USMA officers intend to leave after six years of service and 48 percent intend to leave after ten years of service (Adelsberger, 1988). These data on "intentions" are also supported by data on attrition "behavior", e.g. Hunter (1988) noted that the attrition rates of the 1980 entry-year groups of USMA and ROTC after 4-6 years were 31 percent and 24 percent respectively. The Total Army Personnel Agency notes that, by June 1988 the actual USMA attrition rates of the 1980, 1981, and 1982 year-groups (who come up for retention) were 43 percent, 36 percent and 30 percent respectively (Lakhani, 1988).

RESEARCH APPROACH

We conceptualize that retention intentions of junior officers are directly related to their levels of satisfaction with military life styles (MILSAT) and such other variables as their Retention Military Compensation (RMC) or pay, and their perceived chances of promotion. The RMC is defined to include basic pay plus allowances for quarters and subsistence and the Federal tax advantage of exclusion of these allowances from taxable income. The MILSAT, in turn, is hypothesized to relate to at least ten variables on quality of life in the Army. The first of these predictors is retention intention because the longer an officer stays or intends to stay with the Army, the greater is likely to be his/her MILSAT since development of taste for Army life is a long-term process. Also, longevity puts a "golden handcuff" on the officer by vesting in retirement benefit which is 50 percent of basic pay after 20 years of service. Second, in this era of two income families, an increase in spouse employment is likely to increase MILSAT because spouse employment tends to increase family income and contribute to increase in economic welfare. Third, if a junior officer finds that life in the Army is similar to what was expected prior to joining it, the MILSAT is likely to be higher. Fourth, if an officer perceives that his/her family will be better off if he/she took up a civilian job, then he/she is likely to be dissatisfied with Army

life. Also, if an officer's MILSAT is likely to increase if he/she is satisfied with: (v) in-service educational benefit, (vi) medical benefit, (vii) retirement benefit, (viii) frequency of PCS, (ix) Army's recreational programs, and (x) Army's commissary and Post Exchange (PX) services.

The statistical model to estimate the interdependent relationships of retention intentions and MILSAT is a system of simultaneous equations for each source of commissioning: USMA, ROTC-Regular, ROTC-Scholarship, and officers Directly Appointed (DA) from civilians (e.g. doctors, attorneys and ministers). The required data are developed from Department of Defense (1986), Survey of Officer Personnel, 1985. Since the retention problem is mainly for the junior officers, we defined and sampled them to be those in the grades O-1 to O-4, i.e. the second and first Lieutenants, Captains and Majors. Most of these officers hold these ranks during their first ten years of service. We sorted only the married officers in order to analyze the importance of family factors noted in the recent literature on retention. One of the questions in the Survey asked: "When you finally leave the military, how many total years of service do you expect to have" ? The values of responses to this question were used by us as a dependent variable in the first equation. Another question recorded MILSAT levels of the officers on a seven point Likert scale--from very dissatisfied to very satisfied. These values were used for the dependent variable in the second equation. The values of the predictors of both the equations were also taken from similar Likert scales, except for RMC which was a continuous variable in 1985 dollars, and intended years of service.

RESULTS AND DISCUSSION

Table 1, showing the results of the retention equation, reveals that, as expected, increase in RMC, MILSAT, or the perceived chance of promotion, tends to increase retention intentions of officers from all the sources of commissioning, except for the DA who are not sensitive to RMC. A major reason for the insensitivity of the DA to RMC is that most of them are medical professionals who enter the Army for a short term to obtain expensive training and quit. For example, the Army Finance and Accounting Center (1983) data suggest that the Army's average training cost range from \$103,000 for clinical pathologists (Military Occupational Specialty 61V) to \$385,000 for medical officers (MOS 61Y). The elasticities in Table 1 indicate the percent increase in retention years for a one percent increase in RMC.

The results in Table 1 are used for incremental cost-benefit analysis of retention of USMA versus ROTC officers. The incremental benefit of retention to the Army is the saving in replacement cost. This amount is the cost of education of the USMA or the ROTC officers discussed above. An additional benefit of retention to the Army is the value of increased

performance or productivity of the retained officer because he/she is selected for promotion according to the criterion of "up or out" based on performance. The additional cost of retention is the requirement to increase RMC. Another element of additional cost is the Army's expenditures on longevity cost of promotion and retirement accrual. Since there are no data on the value of improved performance of the retained soldiers, we will assume in this first scenario that this additional benefit is offset by the incremental cost of promotion and retirement.

In Table 2, the elasticities are converted from percents into an increase in the years of intended retention and the increase in RMC dollars. Such a conversion yields the incremental RMC cost per additional year of service of USMA and ROTC officers at \$1,685 and \$7,807 respectively. The RMC cost is higher for ROTC because their elasticity is so low that they need be paid substantially more to obtain an additional year of their service. The incremental benefit to society is the saving in replacement cost, namely, \$225,000 and \$183,000 required to access an USMA and ROTC officer respectively. Since the average intended service of these officers is 20 years, we obtain an average annual value of these benefits by simply dividing the replacement costs by 20, viz. \$11,450 and \$9,365 per year for USMA and ROTC respectively. When these annual values of benefits are divided by the annual values of the RMC cost, viz. \$1,685 and \$7,807, we get the incremental benefit to cost ratio of 6.8 for USMA and 1.2 for ROTC officers. Therefore, we conclude that retention of USMA officers is cost-effective relative to that of ROTC officers. It must be added that retention of ROTC officers is also cost-effective because the incremental benefit exceeds the incremental RMC cost. These benefit-cost ratios did not change significantly when we used the relatively lower estimates (\$191,000 or 175,000) of the replacement costs of USMA officers. In the second scenario, we conservatively assumed that the incremental productivity or benefit of the promoted officers is zero and the only relevant benefit is the saving in replacement cost. Based on the Army Research Institute (1988) Army Manpower Cost System (AMCOS) model, our assumption of promotion of all retained officers from Majors to Lieutenant Colonels resulted in the incremental benefit-cost ratios of 1.5 and 1.2 for USMA and ROTC respectively. Therefore, once again, we conclude that retention of USMA officers is cost-effective relative to that of ROTC officers.

The results of the MILSAT equation revealed (not reported here for brevity) that the predictors common to all the four commissioning groups were the realizations of expectations of military life and retention intentions. The remaining eight variables influenced the four groups differently.

TABLE 1

RESULTS: RETENTION EQ. PREDICTORS

	USMA	ROTC NON-SCHOL.	ROTC SCHOLAR	DA
	(N=209)	(N=643)	(N=372)	(N=228)
RMC:				
COEFF.	14.55*	2.97*	3.38*	0.77NS
ELASTICITY	0.74	0.15	0.18	0.04
MLSATISF	1.48*	1.29*	1.12*	1.07*
PROMOTION CHANCE	0.94*	0.73*	1.03*	0.84*

TABLE 2

SCENARIO 1: INCREMENTAL ANNUAL RMC
COST-BENEFIT TRADEOFFS

	USMA	ROTC
• AVERAGE INTENDED YOS	20	20
- THEREFORE ONE ADDITIONAL YOS =	5%	5%
• AVERAGE RMC PER YEAR =	\$24,500	\$23,500
• RMC ELA.: 10% INCREASE IN PAY INCREASES RETENTION YEARS BY:	7.4%	1.7%
- THEREFORE 5% INCREASE IN YOS WILL REQUIRE RMC INCREASE OF:	7%	30%
- DOLLAR COST OF THIS ANNUAL RMC INCREASE =	\$1,685	\$7,807
• ARMY'S REPLACEMENT COST OF 20 YOS OFFICER =	\$225,000	\$183,000
- THEREFORE ANNUALIZED COST SAVINGS =	\$11,450	\$9,365
• THEREFORE BENEFIT/COST	6.6	1.2

REFERENCES

- Adelsberger, Bernard, J. (1988). West Point: Worth the Cost? Critics, Supporters, Debate Expense of Academy Education. Army Times, May 9, p.12-16.
- Army Finance and Accounting Center (1983). Military Occupational Specialty Training Cost Handbook (MOSB). Fort Benjamin Harrison, Indianapolis.
- Army Research Institute (1988). Army Manpower Cost System Model developed by SRA, Manpower and Personnel Research Lab, Alexandria, VA.
- Department of Defense, Defense Manpower Data Center (1986). 1985 DOD Survey of Officers and Enlisted Personnel: Users Manual and Codebook, Arlington, VA., June 27.
- Doleman, Lt. Gen. Edgar C., USA-Ret. (1985). Abolish West Point? Counterpoint. Armed Forces Journal International, November, 69.
- Gotz, G.A. and J.J. McCall (1983). Sequential Analysis of the Stay/Leave Decision: U.S. Air Force Officers. Management Science. Vol. 29(3), 335-351.
- Hunter, Fumio (1988). Tenure Patterns of U.S. Army Commissioned Officers in the 1970s and 1980s. Technical Report, U.S. Army Research Institute, Alex., VA.
- Johnston, I.D. (1988). Turnover of Junior Army Officers: Master of Science in Management thesis, Naval Post Graduate School, June.
- Lakhani, Hyder (1988). Analysis of Junior Officers Attrition Rates by Race and Sex. Working Paper PUTA-88-10, U.S. Army Research Institute, Alex., Va., August.
- Schemmer Benjamin (1985). Why Waste Money on West Point? The Washington Post, p.C1, December 1; also "Commentary", Armed Forces Journal International, September 1985.
- Scott, Lt. Gen. Williard W., Jr. (1985). One Taxpayer's View of West Point. Armed Forces Journal International, September, p. 81-82
- Smith, David Alton and E. Goon (1987). Spouse Employment and the Retention of Air Force Officers: Some Preliminary Results. Paper presented at the Eastern Economic Society's Annual Meetings, Crystal City, Arlington, VA., March.

A NON-OBTRUSIVE METHOD OF EVALUATING TACTICAL DECISION MAKING IN THE FIELD

Marvin L. Thordsen and Gary A. Klein
Klein Associates

Rex R. Michel and Major Edward W. Sullivan
U.S. Army Research Institute

Background

In the type of group decision making that occurs in military command and control operations it is often possible to measure the quality of the decisions themselves. Criterion measures for the decisions can be developed beforehand and applied to the recorded outputs after the exercise. Measuring the quality of the decision process however is seldom done. The situations are usually too complex and fast moving to permit the level of detailed recording and analysis required for valid process measurement. This paper describes work done by Klein Associates for the Army Research Institute that resulted in the evolutionary development and evaluation of a method for capturing military group decision making process dynamics for use in training feedback and operations analysis.

Over the past several years Klein Associates have performed several basic and applied research projects for the Army Research Institute involving development of methods for extracting expert knowledge. This work is focused on the decision making process as it occurs in real-world situations. The first effort was an investigation of decision making by urban fireground commanders at the scene of a fire (Klein, Calderwood & Clinton-Cirocco, 1986). The method of knowledge elicitation used was a retrospective protocol analysis based on the fireground commanders reconstruction of his step-by-step decisions and commands at a fire scene. Challenging incidents were chosen to increase the probability of recall and to reveal important aspects of expertise not inherent in more routine situations.

The analysis of these interviews lead to the conclusion that these experts performed little conscious analysis of the situation. When faced with a decision point they were able to quickly recognize the situation as some type they were familiar with, recall the typical reaction, perform some brief mental evaluation of the feasibility of that reaction, and then carry it out. They did not appear to be doing any comparisons of different options. The great majority of these decisions were made in a minute or less. Klein has labeled this type of naturalistic decision strategy as Recognition-Primed Decision-making (RPD). It is more fully described in Klein and Calderwood (1986).

Later, Klein Associates observed first-hand team decision making at a large forest fire in Idaho (Taynor, Klein & Thordsen, 1987). This lead to the desire to see if knowledge elicitation could be performed in team decision making situations at the time the decision was made.

The work reported here applied these knowledge elicitation methods to Army command and control decision making. We were interested in seeing not only if the knowledge elicitation methods could be applied to Army command post exercises, but also if the RPD model of decision making found with fireground commanders would also be found with Army tactical decision makers.

We knew that some modification to the data collection would be necessary if we were to collect the data during the exercise. To avoid interrupting the flow of the exercise, interviews with the decision makers would have to be very brief and conducted during breaks or lulls in the action. Klein therefore developed several modifications to their interview technique to test during command post exercises.

There are also differences between the decision making situations encountered by fireground commanders and tactical decision makers that might affect the RPD model. Firefighting decisions are made on the scene and typically have to be made very rapidly. At the level of tactical decision making in which we were interested (i.e., battalion through division), the cues are typically obtained second hand from battlefield reports and summarized information. Also, although time pressures exist, they are less acute at these levels than those facing the fire captain, fewer decisions need to be made instantaneously. A third consideration is that most of the Army decision makers at these levels are familiar with the Military Decision Making Process taught at Army schools. This process involves the deliberate generation, analysis and comparison of alternative courses of action and runs counter to RPD-type decision making. The findings for both the methodology and the model would have implications for the design of decision support systems as well as training.

Three exercises were used for the initial data collection, one each at division, brigade and battalion levels. The last one, at battalion level, proved most productive. It was a well-controlled exercise using the Army Training Battle Simulation System (ATBASS), and the voice recordings with the observer's notes permitted a detailed analysis of the decision making process (Thordsen, Galuska, Klein & Young, 1987). The decision process used by the operations staff was charted in detail for a five hour planning session. We found that planners would evaluate a single concept by gradually examining deeper and deeper branches of the idea for feasibility until it was either accepted, rejected or left hanging due to some distraction. If it was rejected the planners either moved to an entirely different concept or went back up this progressive deepening chain to a point above the flaw and proceeded to follow another branch. Also few such discussions were carried to a final decision, most being interrupted for one reason or another. Only 8% of the transitions were due to resolution of the original topic, while nearly 30% were due to unrelated questions which caused an abrupt change in topic. These data suggest a lack of adequate control over the team decision making process.

Other findings also indicated that a wealth of useful information for both modeling the process and training decision makers can be obtained from such an analysis. The analysis, however, had taken three months to complete. What was needed was a procedure for collecting the required data during the exercise.

such a manner that they could be transcribed directly into the tabular formats and process charts used in the previous analysis. We believed that this would reduce the time for analysis to a few days. We also felt that the knowledge gained from the first analysis and the systematic data collection would permit us to provide feedback to the players on their decision making performance at the end of the exercise. Such rapid feedback was necessary if this performance data collection method was to be an effective training tool.

Method

The vehicle chosen to test the methodology was a week long division-level classroom planning exercise at the Command and General Staff College, Fort Leavenworth. The exercise was part of an experimental class in advanced war-fighting techniques with emphasis on battlefield synchronization and the use of automation to assist planning. The class consisted of 62 students, all of whom were Army majors attending the regular ten month course in staff operations.

The instrument used for data collection was a notational form having columns for the various types of decision making processes uncovered in the analysis of the ARTEASS exercise. Figure 1 shows the format of the data collection form.

DATE _____ LOCATION _____ CODER _____ PAGE _____

WHO	CONTENT/COMMENTS	OPT	INF	DP	APP	SIM	ACT	BRK	ELB	SFT	PRB

Figure 1. Data collection form.

The first column is for the coder to enter his abbreviation for the person speaking or performing the action. The next column allows limited space for a description of the content. No attempt was made to do verbatim recording of conversations. It was feared that too much emphasis on recording content would force the observer to miss the underlying decision processes.

The remaining columns were designed to check the occurrence of ten decision process categories:

OPT (Option Generation): The generation of options and alternatives by the planners.

INF (Information): The introduction of facts.

DE (Decision): The stating of a decision. It is considered a decision if stated by a member of the command group or represents a definite consensus of the group present and is something they have the authority to do.

APP (Appraisal): Any discussion or debate that serves to further the state of the plan but does not introduce new information, make decisions, or generate options.

SIM (Simulation Intrusion): Any time the participant's awareness of the "game" or its apparent artificialities causes the discussion to focus on the simulation per se.

ACT (Action): This refers to activity or a request for activity, usually to gather additional information or detail some aspect of an option.

BRK (Break): Any change in the focus of the planning discussion. The content will contain the cause.

ELB (Situation Assessment Elaboration): Any change in the assessment of the current situation that does not cause a shift in goals. Typically, this results from information that permits a more detailed understanding of what one believed previously.

SFT (Situation Assessment Shift): Any change in the assessment of the current situation that causes a shift in goals.

PRB (Problem): Identification of a potential problem or contradiction, normally occurs in conjunction with appraisal and information processes.

This format assisted the observer in tracking the types of information and level of detail needed for the progressive deepening charts. It also aided in the rapid identification of general patterns and the quantification of results.

During the exercise, a single observer used the data collection form while observing a division level "staff" over several days. A second observer was free to move among other organizations involved and took only general notes. We were seeking to provide rapid feedback to the players concerning examples of good and bad decision processes and the general patterns observed. We also wanted to provide the progressive deepening charts and quantified results to the instructors as soon after the exercise as possible.

Results

We were able to collect the data without interfering with the exercise. There were no formal interviews of the participants. We did occasionally ask someone for clarifying information but such conversations were brief and done during lulls in the action.

We were able to provide feedback at the debriefing immediately following the exercise. This consisted of comments on general patterns of decision making we had observed as well as illustrative incidents. For example we commented that the division staff generally did not aggressively seek needed information from outside sources. We were also able to give specific instances where this affected their performance, e.g., when failure to obtain information from the Corps on the current location of enemy units caused delay in an attack helicopter mission.

Using the data collection form, we were able to flowchart the decision making processes relatively quickly. Charting 20 minute segments of the exercise took about three hours for the first one and about two hours for each additional one. Again we found little concurrent comparison of options but rather the progressive deepening of a single option discussed earlier. Figure 2 shows a five minute segment from these flowcharts.

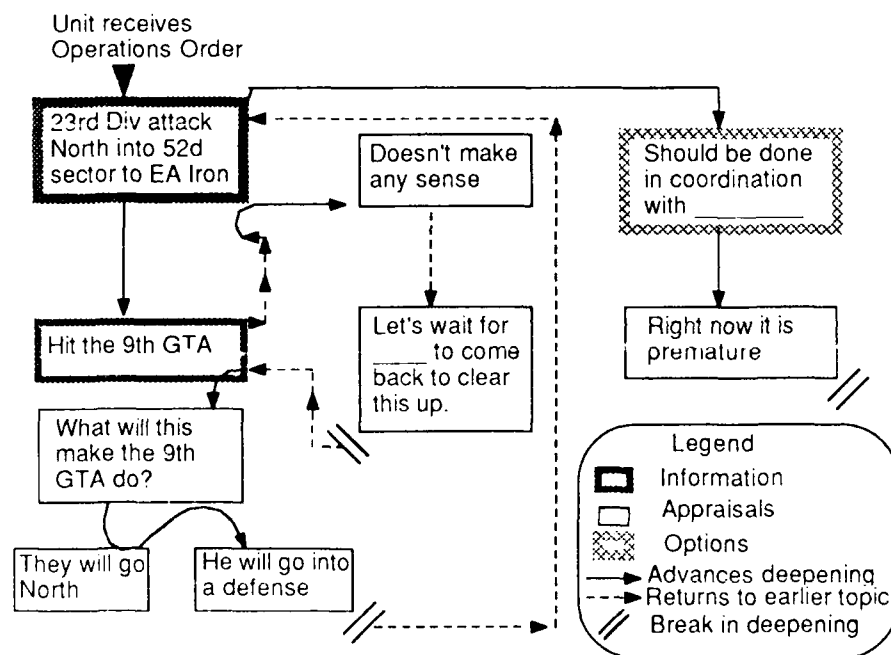


Figure 2. Progressive deepening for a five minute segment of planning.

These flows were not presented to the participants as we were unable to produce them while the exercise was in progress. However, had the exercise occurred other than at the end of the course, we are confident that flows showing their processes during the most critical four hours of play could have been discussed with the class within seven to ten days after the exercise.

Data charts were also produced which showed the amount of participation by each staff member and the frequency of occurrence of the various process categories discussed earlier. The charts took about one day for the analyst to produce. For the six hours and seventeen minutes analyzed in depth, there were sixty changes in topic, most of these unintentional.

Conclusions

We were able to track most of the team processing without interfering with the exercise using our data collection form. The form, when supplemented with general observation techniques, permitted the production of four types of decision making performance feedback: illustrative incidents, general patterns, process flowcharts, and quantifiable data. Although the first two can be available for immediate after action review, the last two types require about seven work days for a single analyst to prepare for a complete eight hour exercise period.

The process flowcharts will allow the players and instructors to review the exercise in detail. They can look at the charts and see the options generated; where, when and why they stopped pursuing them; identify the information input and follow what they did with it. The flows and quantitative data would also be useful for modeling the naturalistic tactical decision making process and in evaluating decision making performance.

A more detailed description of this work is contained in Thordsen and Klein (in press).

References

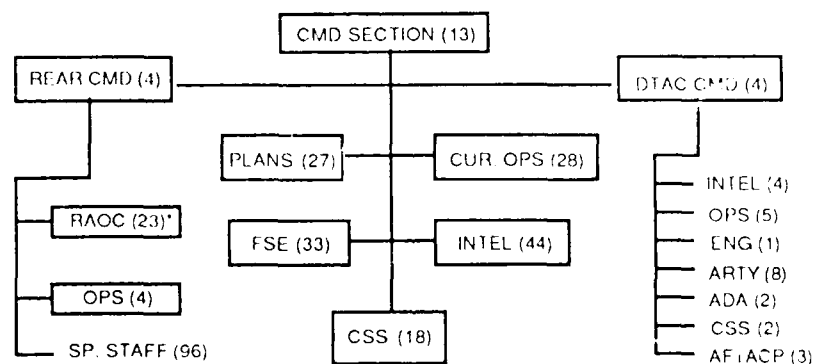
- Klein, G. A., Calderwood, R. & Clinton-Cirocco, A. (1988). Rapid decision making on the fire ground. ARI Technical Report 796. Alexandria, VA: U.S. Army Research Institute.
- Klein, G. A. & MacGregor, D. (1988). Knowledge elicitation of recognition-primed decision making. ARI Technical Report 799. Alexandria, VA: U.S. Army Research Institute.
- Taynor, J., Klein, G. A., & Thordsen, M. (1987). Distributed decision making in wildland firefighting. (KATR-858(A)-04F). Yellow Springs, OH: Klein Associates Inc. Prepared under contract MDA903-85-C-0327 for U.S. Army Research Institute, Alexandria, VA.
- Thordsen, M. L., Galushka, J., Young, S. & Klein, G. A. (1987). Distributed decision making in a command and control planning environment. (KATR-863(C)-87-08F). Yellow Springs, OH: Klein Associates Inc. Prepared under contract MDA903-86-C-0170 for the U.S. Army Research Institute, Alexandria, VA.
- Thordsen, M. L. & Klein, G. A. (In press). Methods for providing direct feedback about decision processes for command and control classroom exercises. Yellow Springs, OH: Klein Associates Inc. Prepared under contract MDA903-86-C-0170 for the U.S. Army Research Institute, Alexandria, VA.

MEASURING DIVISION LEVEL COMMAND AND CONTROL PERFORMANCE

Dr. Lloyd M. Crumley
The United States Army Research Institute
for the Behavioral and Social Sciences
Fort Leavenworth Field Unit

The problems inherent in measuring command and control performance effectiveness at the division command post level are not insignificant. This paper provides an overview and status report of a developing measurement system: the Army Command and Control Evaluation System (ACCES).

Figure 1 shows the layout of a heavy division command post in the field. Even if one disregards the 96 person Special Staff and the 24 persons assigned to Defense Liaison Teams there are still a large number of people performing a broad variety of tasks, tasks which must all be performed, at least adequately, if the battle is to be fought with the minimum of problems. In the command and control world described by FC 101-55 the Division Tactical Operations Center or DTAC is controlled by the Assistant Division Commander Maneuver (ADCM) who with his staff fights the close battle. The Division Rear CP, or DREAR, under the command of the ADC-Support, fights the rear battle and, collocated usually, with the division supply command - or DISCOM - sees that the division is sustained. The division main command post, divided into six functional areas fights the deep battle, monitors the entire battle, and integrates the various battlefield systems into a cohesive whole.



* Also 6 four man defense liaison teams

Figure 1. FC 101-55 heavy division CP layout.

When ARI began its command and control performance measurement work an early step was to review the relevant literature. The literature is rather meager at the division level but overall it provided some help in structuring the problem. The literature demonstrates the efficacy of modelling a division command post in a conceptual framework which enables the researcher to identify those functions, or processes, which need to be performed as the organization attempts to cope with and control a continuously changing environment. The literature also shows that there is a need for great care in determining what

constitutes a suitable command and control effectiveness measure, that care must be taken to assure that the measurement process does not become entangled in the command and control process itself, and that the measurement scheme addresses the soundness of functional mission performance not the correctness of the operational mission performance.

Olmstead, Christensen and Lackey, (1973) in an early command and control research project, articulated a most logical conceptual framework for command and control performance measurement research. In their conceptualization, which drew on the work of management theorists such as E. H. Schein and L. von Bertalanffy, the researchers developed the thesis that "Organizational Effectiveness" (mission accomplishment, productivity, profit, etc) is the final outcome of "Organizational Competence" which they defined as the capacity of the organization to cope with a continuously changing environment. In their model organizational competence was seen as consisting of components that in turn, were composed of basic organizational processes as shown in Figure 2. Their research addressed the question of what is the best measure of command post effectiveness and adopted decision quality as the best measure. Ultimately, after doing what may well be the best battalion research project ever done, Olmstead and his associates reported that the correlation between organizational competence and organizational effectiveness was .93. Very clearly the staff that performs best provides its decision makers with the data they need to make better decisions. Or if, as some writers have said, "the purpose of a staff is to keep the commander from making a mistake" the competent staffs surely do it better than less competent staffs. A fuller discussion of the Olmstead et.al. work is beyond the needs of this paper but it is appropriate to note that it is a report that should be read by anyone concerned with command and control research, training or testing.

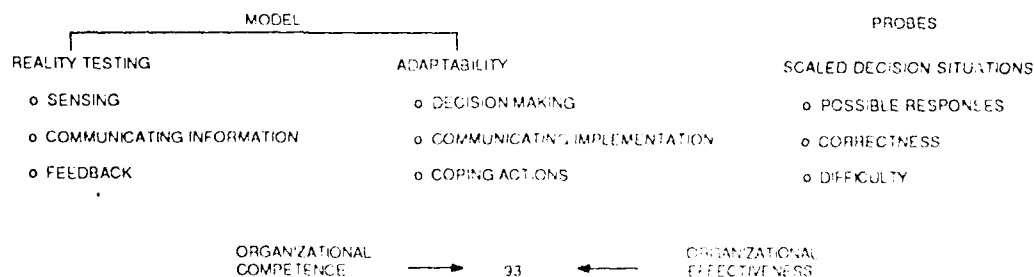


Figure 2. Olmstead et al (1973) research approach.

The literature review also showed that attempts to measure command and control performance have used four different effectiveness measures. Each of the four has both advantages and problems. The already discussed decision quality is obviously a good measure but it was difficult to implement at the battalion level and, quite probably, would be impossible to implement at the division level. A second metric which has been attempted is battle outcome. Battle simulations such as ARTLASS and JHSS provide battle outcome data, and, indeed real - or nearly real - battles such as those at the National Training Center also provide battle outcome data. Unfortunately, while no one doubts

that good command and control will win over bad command and control if all else is equal all else is never equal. Hence, battle outcome data simply does not show significant correlations with staff performance.

ARTEP measures also have significant disadvantages. It is attractive to consider that the trained observer-controller (OC), acting in the role of surrogate supervisor, can check off staff performance on a go-no go list and assign an overall task score. Unfortunately ARTEPS tend to mix functional and operational missions, require skilled persons to make the required judgements, and, since the O-Cs are aware of the battle events, the scores tend to reflect battle outcome.

A fourth organizational effectiveness measure, was proposed by the developers of a command and control measurement technique called the Headquarters Effectiveness Assessment Tool (HEAT). The HEAT paradigm sees the primary command post function as decisions making and develops a rationale to support two effectiveness measures: plan life and degradation mode. In the HEAT concept good plans last longer than bad plans because they can be kept on track by relatively minor changes which do not require a complete decision cycle to accomplish. Good plans also contain elements which show that a greater number of likely alternative battle developments have been considered. These elements manifest themselves by providing a mechanism that enables a headquarters to move to a new plan by going to an alternative which is feasible because some of its special requirements were considered when the old plan was being prepared. This provides a graceful degradation mode that permits getting out a new operations order without going through a complete replanning cycle.

In developing our command and control performance measurement methodology, ACCES, we opted to base our approach on the basic concepts contained in the HEAT research. We consider that a division level measurement scheme is well based if it conforms to the model shown in Figure 3. This approach permits early research to determine how organizational competence affects organizational effectiveness. Later, as data from multiple division experience on the same, or similar scenarios comes available the use of plan life and graceful plan degradation as effectiveness measures can be validated. This later validation is likely to be quite difficult since it would require that battle scenarios be tightly controlled and that the midstream "adjustments" that keep CPXs within acceptable bounds, to enhance training value, be abandoned.

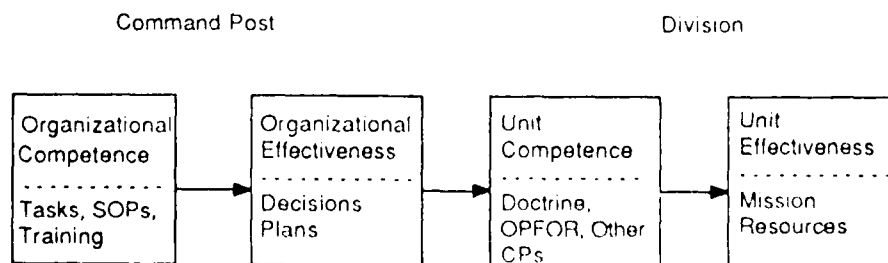


Figure 3. Command post vs. unit commanded model.

The ARI-ACCES model is shown in Figure 4. It has four major concepts embedded in it. First, of course, it depicts a six phase decision cycle. ACCES provides measures for the monitor, assess, generate, plan and direct phases. No measures are provided for the decision block. Instead the number of basic plan assignments changed in the battle plan - or OPCODE - is tallied over the expected life of the plan. The division command posts' assignments in its OPCODE are mission, assets, schedule and boundaries. If changes are many ACCES infers that the planning process is -less effective than if there are few or none. A third set of measures deal with the functions that must be performed in order to keep the decision cycle operating. The tasks, shown as Alpha and Beta tasks in the figure, refer to how well the battle staff passes information throughout the various command post sections and how well the command post maintains its relationship with the exterior world. The former measures relate to coordination, CP network capability and maintaining a common perception of the battle. The latter measures deal with reporting to higher echelons, coordinating with major subordinate commands and the time required to distribute the OPCODE to all critical users.

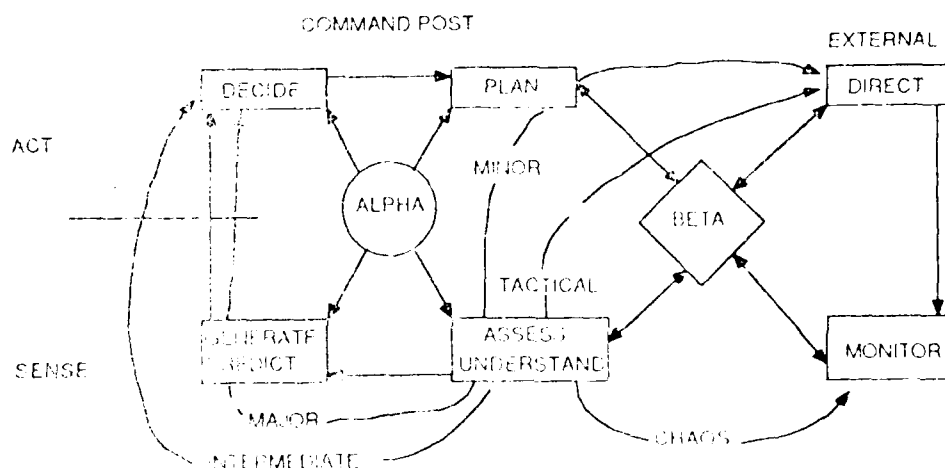


Figure 4. ACCES model showing decision cycle steps, ancillary tasks, and plan degradation levels.

The final concept shown in Figure 4 deals with the magnitude of the deviations, or incursions, that occur as a plan is exposed to the machinations of a thinking enemy. There are five levels of deviations shown. Each is based on how the intellect or perceived future situation differs from the situation the basic plan was devised to create. Tactical and minor changes maintain the original plan. A moderate level incursion occurs when the original plan is deemed to be of little value but a suitable alternative has been considered and is being used. A major incursion requires a full trip through the decision cycle and the fact that the battle staff has been surprised. The chaos state occurs when the situation where the units of the battle have become so obscured that the command post is unable to support a replanning effort and must, therefore, attempt to understand the situation.

These differing levels of adaptive behavior are the basis of the second measure of command and control effectiveness. A commander needs a staff that creates good quality plans that can be kept in effect with only minor adjustments. A commander also needs a staff that identifies likely alternative situations in time to have them considered as the primary plan is developed. If a staff is sufficiently successful it can stay in the minor and intermediate replan portions of the cycle and force the enemy into a situation where he has major replan requirements occurring so often that he ultimately falls into the chaotic situation.

As the HEAT methodology was modified into the division level command and control performance measurement process known as ACCES intermediate versions were tested at division CPXs. The earliest HEAT modification was done on a contract funded by the Operational Test and Evaluation Agency (OTEA). This early version, referred to now as MCS-HEAT, was applied to a division equipped with the Maneuver Control System and to a division that was not so equipped. The purpose of the test was to see if a technology that was designed to measure how competently a division command post performed the processes involved in its functional mission could be used to detect changes in organizational performance that could be attributed to the presence of automated command and control aids. The results of the OTEA program are classified but the two applications did demonstrate that the concept of expressing command and control performance in terms of the processes involved in the conduct of a headquarter's functional missions - such as monitoring, directing, etc. - was a good one.

Three later applications, conducted as part of the ACCES development program, have provided field experience and process information concerning the applicability of the technique. Details of ACCES developments are beyond the scope of this presentation but some changes are worth noting. As ACCES evolved from HEAT the model changed, the number of decision routes as a function of incongruence level went from three to five, there were changes in the organizational competence measures, and the primary organizational effectiveness measure became number of plan assignment changes rather than plan life.

The experience gained has made it possible to make some inferences concerning the value and application of ACCES. We have conducted a division application with only seven observer teams to see if we could obtain results with coverage at those points where system integration of division assets was being done as opposed to twelve team applications where data were also obtained at subsystem integrative points. Results showed that data from only the G-3 controlled, system integration points provided an adequate assessment of command and control but that subsystem integration observations were needed if training feedback was to be most effective. It may be of interest to note that in a three or four day CPX there are about 150 decisions when both system and subsystem integration decisions are collected. Of these, approximately 90 are made at system integration points (DTAC, Current Operations, Plans, and DREAR).

A sample of ACCES results are shown in Table 1. The data, from an exercise shows a few of the 43 ACCES measures. Scores reflect a division command group who, due to unforeseen circumstances, found DTAC, DMAIN Plans and DMAIN Current Operations trying to solve problems that would normally have been left

to brigade command posts. As the staff became more and more involved in what could have become a chaotic situation they got more and more information about friendly units who they could contact and gradually lost control of the processes of maintaining good communications with their own units who were busy responding to their own problems.

Table 1

Some examples of ACCES data

Measure	Day 1	Day 2	Day 3
Understanding time	30	27	6
Understanding quality			
OPFOR	79	75	71
Self	75	95	89
Unit location			
OPFOR	76	77	70
Self	59	59	40

The ACCES program has been quite successful to date. In the near future we feel it may be possible to reduce the time required to complete the observer notes - to data sheets - to analysis cycle time and, perhaps, to do some real time data reduction. Some work done in cooperation with the Battle Command Training Program suggests that the goal of reducing the data collection and analysis parts of the cycle can be met. Real time data reduction which can provide ACCES measures for After Action Reviews (AARs) appears much more difficult. It does seem, however, that the presence of observers 24 hours per day creates a significant body of data which is available in real time for providing training feedback in AARs as the CPX is proceeding. It is also clear that application of a standardized methodology which quantifies command and control performance can provide a good tool for future research in the area.

REFERENCES

- Defense Systems, Inc. (1984). The headquarters effectiveness assessment tool: Users Manual. McLean, VA.
- Olmstead, J. A., Christensen, H. E., and Lackey, L. L. (1973). Components of organizational competence: test of a conceptual framework. HUMRRO TR-73-19. Alexandria, VA: Human Resources Research Organization.

SIMULATION-BASED C³ TESTING FOR ARMOR PLATOON LEADERS

Robert S. Du Bois
Universal Energy Systems, Inc.
Carl W. Lickteig
U.S. Army Research Institute

Effective small-unit, crew and platoon, command, control, and communication (C³) is central to the success of the Armor Maneuver Force on the future battlefield. Current armor doctrine stresses speed, aggressiveness, deception, penetration, and synchronization (Department of the Army, 1986; U.S. Army Armor Center, 1982). Tank crew and platoon commanders at the forward edge of the battle area (FEBA) must see the whole battlefield and react rapidly and flawlessly to deceive and destroy a numerically superior force. These commanders must recognize battlefield changes, accurately and quickly report these changes, and rapidly receive and correctly interpret reports and execute orders from other commanders. These C³ requirements will be even more critical as the Army introduces increasingly sophisticated computer-based C³ capabilities, like the Battlefield Management System (Blasche and Lickteig, 1984; Lickteig, 1987).

Preparing small-unit armor commanders for the C³ requirements of the future battlefield involves providing these commanders sufficient opportunities for practice and feedback under realistic task conditions. Armor evaluators are currently challenged to reliably assess C³ skills in the context of combat mission field exercises, like those conducted at the Army's National Training Center. While these exercises offer the advantage of realism, C³ assessments are frequently affected by aspects of the battle not directly related to C³, such as doctrine, mission success, and gunnery (Crumley, 1988; Wheaton and Boycan, 1982). The complexity of coordinating multi-combat vehicle exercises, and the time and resources required for field testing, often force test developers to include an inadequate number of C³ tasks and less than optimal measurement methods, such as the ratings of observers stationed, off-tank, at selected points along the battlefield.

A promising approach for at least partially overcoming these C³ assessment problems is to use interactive combat simulation. Simulation systems, used frequently for C³ assessments at the armor battalion-level and higher, have rarely been used for small-unit, crew and platoon assessments (Crumley, 1988).

One of the Army's most advanced simulation test beds, Simulation Networking - Developmental (SIMNET-D), interactively links a variety of combined arms, soldier-in-the-loop, simulators (Bolt, Beranek and Newman (BBN) Laboratories, 1986; Miller and Chung, 1987; Perceptronics, 1986). Images generated provide crew members real-time updates of the terrain features, other vehicles and weapon effects within a 3,500 meter radius while moving, scanning or shooting on a 50 km by 75 km battlefield. A sound system recreates realistic battlefield acoustics, such as weapons firing and track movement appropriate to tank speeds, terrain surface, steering and gear changes. A semi-automated forces simulation system allows researchers to use computer-generated opposing and friendly forces (OPFOR and BLUEFOR). Planned SIMNET enhancements include JETNET and AIRNET, which add Air Force fixed-wing aircraft and Army helicopters to the SIMNET battlefield. Future enhancements may extend SIMNET architecture to include Navy and Marine vehicles and support stations. Armor evaluators have suggested that SIMNET provides an excellent test bed for evaluating small-unit C³ proficiency (Gound and Schwab, 1988).

A goal of this Army Research Institute (ARI) research program in C³ is to capitalize on SIMNET's capability to rapidly, repeatedly, and realistically generate C³ assessment exercises. SIMNET provides an environment for creating a series of standardized battlefield conditions, including both friendly and enemy situations, for initiating C³ tasks. Furthermore, SIMNET provides performance measurement capabilities supporting the collection of objective C³ measures too costly or dangerous to accurately capture in the field. By indicating how small-unit armor commanders can perform critical C³ tasks, these C³ assessment exercises will allow armor trainers and evaluators to assess C³ skills and diagnose training deficiencies. This paper describes the initial test development stages of this program, as well as evaluation efforts in-progress and planned for the future.

Test Development

Identifying small-unit C³ tasks

The first step in developing small-unit armor commander C³ assessment exercises was to identify and review the domain of crew and platoon commander C³ performance objectives. The goal of this step was to identify the C³ performance requirements of armor crew and platoon commanders which can be rapidly, reliably, and realistically initiated and objectively measured in SIMNET. A cooperative effort between researchers and four Armor subject matter experts (SMEs), and previous Army C³ task analyses (e.g., Department of the Army, 1987; U.S. Army Armor School, 1988), helped in this effort. The Armor SMEs included an instructor for the armor platoon leader's basic course (AOB), an Armor test officer, and two platoon leaders from active armor units.

Overall, several armor crew and platoon C³ tasks were identified for assessment in SIMNET. These tasks include: preparing, executing, receiving, and issuing fragmentary orders; bypassing obstacles; reporting enemy indirect fire; reporting own location; preparing, sending, and receiving reports of enemy activity; reacting to enemy direct and indirect fire; calling for, and adjusting, indirect fire; and selecting and occupying battle positions.

Generating Test Exercises

A goal of the C³ test exercises is to require small-unit armor commanders to complete critical C³ tasks in standardized and realistic, yet directly measurable and discrete, combat events. Hence, a multi-dimensional testing approach is planned for assessing crew and platoon commander C³ proficiency. This approach involves the development of three distinct types of test exercises: (1) vignettes or combat items, (2) crew tactical exercises, and (3) platoon combat mission scenarios. Each of these types of exercises are specifically designed to vary on the level of standardization and realism they provide. While each of these test approaches is described below, only the crew tactical exercises and platoon combat mission scenarios have currently been developed.

Vignettes. Combat items, vignettes, represent an attempt to use simulation to assess the individual C³ proficiency of crew and platoon commanders using discrete C³ task items. These vignettes allow multiple crew and platoon commanders, located at identical battlefield locations but unable to hear or see each other, to observe and react to standardized sets of visual and auditory stimuli prompting the performance of C³ tasks. Vignettes are the simulation-based equivalent of a paper-and-pencil C³ test. Several commanders are simultaneously presented with the same set of stimuli cues, combat events, across varying conditions. However, while vignettes support objective measurement and standardization, they are not as realistic as collective combat missions. The SIMNET-D test bed is currently being configured to support vignette-based C³ assessment exercises.

Crew Tactical Exercises. Crew tactical exercises represent an extension of the vignette approach designed to improve the realism of the C³ task requirements. Rather than rapidly placing commanders in stationary combat situations on the SIMNET battlefield, the crew tactical exercises are designed to place crew and platoon commanders in more fluid single-tank tactical exercises. Commanders are given a road and cross country route to negotiate. Along this route, each commander is presented with selected visual and auditory stimuli prompting critical C³ tasks. Hence, the context of managing a crew and mission in a moving tank and not being completely aware of one's own location is maintained in these exercises. The increased realism of crew tactical exercises over vignettes does not come without some problems, however. Crew tactical exercises can only be administered to a single commander at a time and control over stimulus conditions is reduced. While

similar to single tank field tactical exercises used in Army tank commander and platoon leader basic courses, these SIMNET-based exercises include a variety of stimulus and feedback conditions which cannot be included in field courses for cost or safety reasons.

Platoon Combat Mission Scenarios. A third set of exercises represents the simulation-based equivalent of a combat mission field exercise. Crew and platoon commander C³ proficiency is assessed in the fluid, tactical environment of a platoon combat mission scenario. This scenario includes both offensive and defensive armor missions, as well as frequent mission changes resulting from fragmentary orders. Commanders must now complete C³ tasks while not only managing their crews in a mission environment, but also coordinating their actions with other friendly commanders in the platoon. While realism is improved, evaluators are now challenged to isolate individual commander C³ proficiency from a collective and complex mission exercise. Each individual platoon may execute the same mission with significantly different approaches. Moreover, combat missions require not only effective C³, but also gunnery and maneuver skills.

Du Bois and Smith (in press) demonstrated the ability of SIMNET to support effective platoon combat mission testing in an evaluation of the armor position navigation (POSNAV) display.

Development of the crew tactical exercises and platoon combat mission scenarios was completed in three phases. First, draft exercises were developed based on the judgments of the four SMEs who identified the C³ tasks. Exercises currently used by the Army's Armor School served as models during this phase for some tasks. Second, the exercises were pretested. Three Armor platoons from active armor units at Fort Knox, KY, repeatedly completed each draft exercise. One platoon was evaluated each week across three weeks of pilot testing. Third, soldier and test controller reactions to these exercises, and initial data analyses, resulted in some modifications to these exercises. For example, some stimuli had to be revised to ensure their salience for initiating C³ tasks.

Criterion Measures

SIMNET-D provides capabilities which allow the collection of critical C³ measures too costly or dangerous to separately gather in the field. In support of SIMNET's distributive network, each combat simulator and an exercise Management, Command and Control (MCC) system continually broadcast information related to vehicle appearance and status, direct and indirect firing events, vehicle collisions, and service requests and receipts. Additionally, a Plan View Display (PVD) allows evaluators to view a combat exercise as seen from a "bird's eye view" above the simulated battlefield in real-time or playback. A flagging function linked to the PVD allows researchers to time-stamp selected events for later analysis. A time-synchronized multi-terminal display system allows

the collection of radio traffic data critical for C³ assessments. On-tank data collectors can be used to collect critical behavioral measures within each tank on the battlefield. The current C³ research program exploits these SIMNET data collection capabilities.

For example, a critical C³ requirement of small-unit commanders is the preparation and transmission of battlefield reports, including reports of enemy vehicle locations (contact and spot reports), reports of bombings (shell reports), requests for indirect fire support (call for fire and adjust reports), and various status reports (situation reports, logistics reports).

On the future battlefield, the speed and accuracy with which reports are prepared and forwarded will be critical to mission success. SIMNET's MCC packets allow us to verify the accuracy of these reports by comparing reported locations with actual locations. The PVD flagging function, voice recorder, and on-board data collectors also allow the collection and analysis of other critical C³ measures. These measures include the speed with which reports follow their respective battlefield stimuli, the appropriateness of each report's format and content, and the appropriateness of the commander's behavior in preparing and sending the report. In the field, evaluators are often not able to collect critical C³ measures accurately. Now, researchers must choose between the numerous criterion measures available with SIMNET. This ARI research program will help in determining the reliability and validity of alternative C³ criterion measures.

Current Evaluation Efforts

Currently forty-eight tank crews, twelve platoons, are participating in an evaluation of two crew tactical exercises and three platoon combat mission scenarios. A four-day test schedule per platoon includes a day and a half of SIMNET-based training, a four-hour crew tactical exercise, and two four-hour platoon combat missions. The data from this evaluation will allow us to examine the psychometric properties of the two types of C³ assessment exercises currently developed. Both reliability and validity of the exercises are being evaluated.

These results will provide the impetus for exercise improvements and future investigations aimed at developing a complete set of C³ assessment exercises for use by Army small-unit evaluators and trainers. Future efforts will also evaluate the training utility of these exercises. Lessons learned from this research will promote the effective development of SIMNET-based exercises for other domains, including C³ assessments of larger armor units, and provide reliable and valid instruments which can generate criteria for soldier-in-the-loop evaluations of new armor systems.

References

- Blasche, T. R. and Lickteig, C. W. (1984). Utilization of a vehicle integrated intelligence [V(INT)2] system in Armor units. ARI Technical Report 1374. Fort Knox, KY: U. S. Army Research Institute for the Behavioral and Social Sciences.
- Bolt, Beranek, and Newman Laboratories Inc. (1986). Developmental SIMNET data handbook [Draft]. Cambridge, MA: Author.
- Crumley, L. M. (1988). Review of research and methodologies relevant to Army command and control performance requirements. DRAFT Technical Report. Fort Leavenworth, KA: U. S. Army Research Institute for the Behavioral and Social Sciences.
- Department of the Army. (1986). Field manual 100-5: Operations. Washington, DC: Headquarters.
- Du Bois, R. S. and Smith, P. G. (in press). The effect of the position navigation (POSNAV) system on the performance of armor crews and platoons. Technical Report. Fort Knox, KY: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Gound, D. and Schwab J. (1988). Concept evaluation program of Simulation Networking (SIMNET). AEB Technical Report 86-CEP-0345. Department of the Army. Fort Knox, KY: U.S. Armor and Engineer Board.
- Lickteig, C. W. (1987). Automated command, control, and communication in lower echelon armor units. In Proceedings of the 29th Annual Conference of the Military Testing Association. Ottawa, Canada: Military Testing Association.
- Miller, D. C. and Chung, J. W. (1987). SIMNET-D capabilities and overview. Cambridge, MA: Bolt, Beranek, and Newman Laboratories, Inc.
- Perceptronics, Inc. (1986). SIMNET master documentation [Draft]. Woodland Hills, CA: Training and Simulation vision.
- U.S. Army Armor Center. (1982). The airland battle: A collection of readings from professional journals. Washington, DC: U.S. Army Armor School.
- Wheaton, G. R. and Boycan, G. G. (1982). Methods of evaluating tank platoon battle run performance: A perspective. Technical Report 457. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

ARMY COMMISSIONED AND NONCOMMISSIONED OFFICER LEADER REQUIREMENTS¹

Alma G. Steinberg and Julia A. Leaman
U.S. Army Research Institute

Background

The purpose of this paper is to use empirical data to identify the components of Army leadership and to demonstrate that within the area of Army leadership the actual role of Army leaders is related to assignment characteristics such as type of officer, branch, type of organizational unit, location. The empirical data are the responses of 10,978 Army commissioned and noncommissioned officers to the Army Leader Requirements Survey.

The Army Leader Requirements Survey is a task analysis that addresses the leadership portion of the job of commissioned officers and noncommissioned officers (NCO). The task analysis was conducted in order to provide the Army's leadership doctrine and curricula developers with empirical data on which to refine their multi-level leadership training and education programs. It was developed from iterative interviews with more than 200 officers and NCO and then reviewed by subject matter experts at the Center for Army Leadership and the U.S. Army Sergeants Major Academy for clarity, accuracy, organization, and completeness. The survey instrument contains 560 leadership tasks divided into 20 individual duty areas (Steinberg, 1987). Incumbents were asked to respond to only those tasks which they perform in their current duty assignment.

The survey respondents were 5033 commissioned officers in grades O1 through O6 and 5945 NCO in grades E5 through E9. The respondents were distributed fairly evenly across all branches (i.e., the percentage of commissioned officers and NCO survey respondents by branch ranged from 2.5% to 9.8%). The respondents were from two types of organizational units, Table of Organization and Equipment (TOE) and Table of Distribution and Allowances (TDA). TOE organizations are structured for a wartime mission and are based on a 24-hour day whereas TDA organizations are peacetime oriented and are based on a work day of 8 hours (Headquarters Department of the Army, 1987). For the commissioned officer respondents, 1563 were in TOE units and 3161 were in TDA units. For the NCO respondents, 2657 were in TOE units and 2468 were in TDA units. For both the commissioned and noncommissioned officers, almost two-thirds were located within the continental United States (CONUS) and the remainder were located outside the continental United States (OCONUS).

¹The views expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Institute or the Department of the Army.

Results

First, the components of the concept of Army leadership are presented and then examples are provided to demonstrate that within the area of Army leadership, leaders differ in what they do, depending on assignment characteristics. The assignment characteristics considered here are type of officer (commissioned or noncommissioned), branch, type of organizational unit, and location. In each case, the examples are based upon leadership tasks performed by most respondents in the group under consideration. "Most" is operationally defined as 66.6% or more of the respondents indicating that they performed the task. For each group of respondents examined in this paper, 90% confidence intervals were calculated around the 66.6% performing cutoff point. Then the tasks that were above the lower confidence interval cutoff point were selected.

Components of Army Leadership

The 26 individual duty areas empirically derived for the Leader Requirements Survey can be grouped into four global duty areas circumscribing Army leadership: (A) Train, Teach, and Develop; (B) Motivate; (C) Resource; and (D) Provide Direction (see Table 1). These four areas follow from the definition of Army leadership which states that leadership is the "process by which a soldier influences others to accomplish the mission" (Headquarters, Department of the Army, 1983). In order to influence others to accomplish the mission one needs to: (a) train, teach, and develop them so that they can do what is necessary to accomplish the mission; (b) motivate them so that they will do what is required; (c) provide the resources for them to do what is required; (d) provide direction so that they know what to do. Note that all these components need to be performed in order to get others to accomplish the mission, but that they need not be done by the leader alone.

Leadership Tasks as a Function of Type of Officer

Table 1 provides the number of leadership tasks performed by most commissioned and officers and by most NCO. Most of the commissioned officers reported performing 32 of the 560 leadership tasks and most of the NCO reported performing 47. These tasks were distributed differently for commissioned and noncommissioned officers. More tasks fell in the first two global duty areas "Train, Teach, and Develop" and "Motivate" for the NCO than for the commissioned officers. The reverse pattern held for the remaining two global duty areas. More tasks fell in the "Resource" and "Provide Direction" areas for the commissioned officers than for the NCO. The global duty area with the most tasks for NCO was "Train, Teach, and Develop." For commissioned officers, it was "Provide Direction."

For each of the individual duty areas for commissioned officers and NCO, the tasks performed by most of one group are

Table 1

	Number of Tasks Performed by Most			
	Officers n=5033	NCO n=5945	Infantry Officers n=248	Chaplain Officers n=258 NCO n=155
<u>Global and Individual Duty Areas</u>				
GLOBAL DUTY A: Train, Teach, and Develop				
A. Train soldiers (21)*	2	14	10	7
B. Teach soldiers (18)	0	5	5	6
C. Develop leaders (21)	4	10	12	13
D. Plan and conduct training (42)	0	1	4	1
E. Train in the field to enter combat (44)	0	0	0	0
TOTAL:	6	30	31	27
GLOBAL DUTY B: Motivate				
F. Motivate others (the what) (13)	1	1	3	3
G. Motivate others (the how) (42)	2	6	13	11
H. Develop unit cohesion (52)	0	0	6	11
I. Reward and discipline subordinates (30)	1	1	5	6
J. Take care of soldiers (33)	1	2	3	7
TOTAL:	5	10	30	38
GLOBAL DUTY C: Resource				
K. Manage resources (40)	9	2	11	10
TOTAL:	9	2	11	10
GLOBAL DUTY D: Provide Direction				
L. Perform/supervise admin. functions (26)	1	0	3	3
M. Coordinate with others outside the unit (20)	0	0	0	0
N. Supervise others (20)	2	2	2	2
O. Maintain 2-way info. exchange with subordinates (21)	6	0	10	7
P. Maintain 2-way info. exchange with superiors (17)	3	0	5	5
Q. Monitor and evaluate performance (38)	0	0	5	3
R. Conduct counseling (24)	0	3	3	2
S. Establish direction of your unit/element (13)	0	0	9	12
T. Provide input for direction of larger org. (25)	0	0	0	0
TOTAL:	12	5	37	32
GRAND TOTAL:	32	47	109	107
*Total number of tasks in each duty area.				

all a subset of the tasks performed by most of the other group. Thus, for example, in individual duty "Train Soldiers", the two tasks performed by most commissioned officers are among the 14 tasks in this individual duty area performed by most of the NCO (see Table 1).

In the area of "Train, Teach, and Develop," the common tasks performed by most commissioned officers and most NCO were general in nature. They were: (a) improve the performance of subordinates, (b) develop good work habits in soldiers, (c) delegate decision-making to subordinates, (d) delegate authority to the lowest appropriate level, (e) train subordinates to take initiative, and (f) support decisions of subordinate leaders. The tasks in this area performed by most NCO, but not by most commissioned officers, tended to be more specific and related to: (a) the job of the enlisted soldier and skills required to perform it, and (b) individual development. NCO reported that they train soldiers to be technically and tactically proficient, to do their jobs without supervision, to perform common tasks and basic military skills, and to pass skill qualification tests (SQT). They also reported that they teach soldiers to perform maintenance regularly, check their own work, solve problems, and meet time requirements. In the area of individual development, they indicated that they teach the soldiers personal discipline, handling of stress, recognition of the ethical dimensions of their decisions and behaviors, and leadership. They also reported that they allow subordinate leaders to learn from their mistakes, recommend military training and civilian education, and provide soldiers the opportunity to receive formal training.

In the area of "Motivate," the common tasks most commissioned officers and most NCO reported performing were: (a) motivate subordinates, (b) set the example, (c) demonstrate Army values, (d) tell soldiers when they are performing well, and (e) promote physical fitness. The tasks performed in this area by most NCO, but not by most commissioned officers, had a more personal aspect to them. They involved (a) remaining available to immediate subordinates until they finish for the day, (b) explaining why tasks need to be done, (c) having face-to-face contact with immediate subordinates on a daily basis, and (d) assisting subordinates with their personal problems.

In the "Resource" area, the tasks that both most commissioned officers and most NCO reported performing were: (a) manage time, and (b) seek ways to improve productivity. Most commissioned officers, but not NCO, reported managing other resources in addition to time. These included people/manpower, information, and things (money, supplies, equipment, etc.). They also indicated that they (a) conduct crisis management (put out fires), (b) solve each problem in order of priority, (c) decide on changes in scheduled activities, and (d) determine their own responsibilities.

The tasks common to both most commissioned officers and most NCO in the "Provide Direction" area were: (a) supervise U.S. soldiers, and (b) supervise male soldiers. The 10 additional tasks that commissioned officers, but not the NCO, reported performing involved communication. One task was "edit and proofread written materials" and the remaining nine involved maintaining two-way communication with subordinates and with superiors. Examples include (a) establish communication channels, (b) encourage upward and downward communication, (c) explain the "why" of things to higher-ranked individuals. In this area of "Provide Direction", there also were three counseling tasks that were performed by most NCO but not by most commissioned officers. These tasks were: (a) counsel male soldiers on their performance, (b) write counseling statements, and (c) make on-the-spot corrections.

Leadership Role as a Function of Branch

When branch is taken into account, interesting differences surface. For example, many more tasks were performed by most infantry and most chaplain commissioned officers and NCO than for most commissioned officers and NCO regardless of branch (Table 1). The individual duties that gained the most tasks for commissioned officers for both branches were Develop Leaders and Motivate Others (The How); for infantry alone, Train Soldiers and Establish Direction of your Unit/Element; for chaplain alone, Motivate Others (The What) and Conduct Counseling. The individual duties that gained the most tasks for infantry NCO were Develop Leaders, Motivate Others (The How), Develop Unit Cohesion, Reward and Discipline Subordinates, and Take Care of Soldiers. On the other hand, for chaplain NCO they were Plan and Conduct Training, Reward and Discipline Subordinates, and Maintain 2-Way Information Exchange with Subordinates.

Leadership Role as a Function of Type of Organization

Many more leadership tasks were performed by most commissioned and noncommissioned officers in TOE units than in TDA units. Most commissioned officers in TOE units performed 137 tasks and in TDA units 26. Most NCO in TOE units performed 103 tasks and in TDA units 31. For both commissioned and noncommissioned officers, the global duties with the most gain for TOE units over TDA units were Train, Teach, and Develop, and Motivate; the global duty with the least change was Resource. With respect to individual duties, for commissioned officers, those with a gain of nine or more tasks for TOE over TDA were Train Soldiers, Develop Leaders, Motivate Others (The How), and Establish Direction of Your Unit/Element. For NCO, they were Develop Leaders and Motivate Others (The How).

Leadership Role as a Function of Location

More leadership tasks were performed by most commissioned and noncommissioned officers OCONUS than CONUS. Most

commissioned officers in OCONUS performed 76 tasks and in CONUS 28. Most NCO in OCONUS performed 86 tasks and in CONUS 41. For both commissioned and noncommissioned officers the global duties with the most gain for OCONUS over CONUS were Train, Teach, and Develop and Motivate; the global duty with the least change was Resource. With respect to individual duties, for commissioned officers and NCO, the only duty area with a gain of nine or more tasks for OCONUS over CONUS was Motivate Others (The How).

Conclusions

It has been shown that (a) four components of Army leadership are Train, Teach, and Develop; Motivate; Resource; Provide Direction; and (b) within the area of Army leadership, leader roles are related to characteristics of assignments such as type of officer, branch, type of organizational unit, and location. In previous publications, it also has been shown that rank/grade is a factor (Steinberg and Leaman, 1988a; Steinberg and Leaman, 1988b). The results presented here have implications for the doctrine writer, the designer of leadership training and education programs, the leadership performance evaluator, and the leadership researcher. The first important step for each of these should be to clearly define the group of leaders their work addresses. It is important for doctrine, training, and education to reflect the differences in leader requirements that are deemed desirable for various groups and to correct the differences that are not. For researchers, the results demonstrate the importance of appropriate sampling to represent the leadership population of interest, selection of content to study appropriate to the group under consideration, and clarification of the group to which their research conclusions apply.

References

- Headquarters Department of the Army (1983). Military leadership, FM 22-100. Washington, DC.
- Headquarters Department of the Army (1987). Manpower requirements criteria (MARC) - Tables of organization and equipment (Army Regulation 570-2). Washington, DC.
- Steinberg, A. G. (1987). The leader requirements survey package (ARI Research Product #87-21). Alexandria, VA: U.S. Army Research Institute.
- Steinberg, A. G., & Leaman, J. A. (1988a). The Army leader requirements task analysis: Preliminary commissioned officer results, (LMTA Working Paper #88-03). Alexandria, VA: U.S. Army Research Institute.
- Steinberg, A. G., & Leaman, J. A. (1988b). The Army leader requirements task analysis: Noncommissioned officer results, (LMTA Working Paper #88-07). Alexandria, VA: U.S. Army Research Institute.

WHAT DO ARMY SERGEANTS MAJOR IN STAFF POSITIONS DO?

Gilbert L. Neal, James D. Moreland, John E. LaVerne
and F. Edward Saia

U.S. Army TRADOC Analysis Command-White Sands Missile Range

Most studies of noncommissioned officers (NCO) deal with the training or job of the junior NCO. This paper, based on Moreland et al. (1987), addresses the training and the job of the sergeant(s) major (SGM), the top Army NCO rank.

The 22-week Sergeants Major Course (SMC) at the U.S. Army Sergeants Major Academy (USASMA), Fort Bliss, Texas prepares Department of the Army (DA) selected master sergeants (MSG) and first sergeants (1SG) for sergeant major (SGM) and command sergeant major (CSM) duty positions when they are promoted to pay grade E-9. The SMC is the capstone course of the Army's Noncommissioned Officer Education System (NCOES) (DA, 1984). It prepares graduates for positions of responsibility within DA and the Department of Defense (DOD). The SMC program of instruction (USASMA, 1985) contains the following: (1) military studies; (2) national security affairs; (3) leadership; (4) resource management; and, (5) the professional development program.

The Army SGM strength was 4515 in the 1986-87 time frame. Approximately 29% of these SGM were in the CSM program in Military Occupational Specialty (MOS) 00Z. The majority of the SGM (71%) serve in staff-type duty positions, representing 64 different career MOSs and 33 different career management fields (CMF). MOS densities ranged from 543 (MOS 11B50, Infantry Senior Sergeant) to one (e.g., MOS 35U50, Biomedical Equipment Maintenance Chief). SGM duty positions are found at all levels of command from battalion and post through DA and DOD, to include international commands. This study focused on SGM in staff positions, not CSMs. The former are heterogeneous in MOS, duty assignments, and levels of assignments.

This study was carried out to assist USASMA determine how well the SMC was preparing graduates for staff assignments. This was to be accomplished assessing (1) what SGM do on the job, (2) job skill and knowledge requirements, (3) professional development and training needs, and (4) perceived effectiveness of NCO common leader training.

METHOD

Approach and Rationale. To accomplish study objectives, the SGM's job was examined from the perspectives of both the incumbent SGM and the immediate supervisors of SGM. Since SGM and their supervisors were in senior level

The findings of this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents. Opinions presented are those of the author(s) and don't represent those of the Department of the Army.

assignments, the structured interview was used as the method of data collection. It was decided that the interview would receive high acceptance from senior level personnel and would provide the best rapport for data collection.

Data Collection Instruments.

Two parallel structured interview guides were developed--one for SGM and one for supervisors. The SGM's guide addressed the following areas from the SGM's viewpoint: (1) assigned organization's mission; (2) tasks, duties, and responsibilities performed in current job; (3) task, etc., that consumes most time; (4) most important task, etc., performed; (5) most difficult task, etc., to learn; (6) how job would change in wartime; (7) how current job differs from last SGM assignment; (8) formal training that helped perform current job; (9) formal training that would have helped perform current job; (10) past assignments that provided experience to do current job; (11) past experiences that would have helped perform current job; (12) supervisor's perception of the SGM's role and job; (13) how effectively the supervisor uses the SGM; (14) how supervisor could be oriented to utilize a SGM; (15) training needed to work with other U.S. services; (16) training needed to work with civilian personnel; (17) how SMC training has helped perform job; (18) what should be emphasized in the future SMC; (19) how other SGM perform their job; and, (20) status of SGM's job description.

The supervisor's guide paralleled the SGM guide, except questions were worded from the supervisor's viewpoint; certain SGM-unique questions were deleted; and certain supervisor unique questions were added. The latter concerned: (1) what SGM couldn't do because of lack of experience or training; (2) strengths and weaknesses of SGM; (3) skills and knowledges of other service SGM, if observed; (4) how the SMC should prepare SGM for the job; (5) what a supervisor needs to know about SGM training and experience; and, (6) status of incumbent SGM's job description.

In addition, a fill-in-the-blank military experience and training background questionnaire for SGM was developed. Supervisor background information was collected as part of the interview. Questions dealt with rank and experience as a supervisor of SGM.

The interview guides and questionnaires were pretested at White Sands Missile Range, NM and at Fort Bliss, TX. Modifications were made in each prior to data collection.

Data Collection Procedures.

The interviews were conducted by a team of trained interviewers consisting of two civilians (grades GS-13 and 14), three commissioned officers (two LTC and one CPT), and three SGM. The SGM interviewed SGM; the officers interviewed supervisors; and, the civilians interviewed both SGM and supervisors.

Interviews were conducted in locations that assured privacy and confidentiality. Interviewee answers were recorded on the interview guides by the interviewer. SGM interviews were one and one-half hours since SGM had to complete a background questionnaire. Supervisor interviews took approximately one hour.

Sampling and Design.

This study was designed to identify common tasks, duties and responsibilities and associated skills knowledges and knowledges across MOSs, duty assignments, organizations, and levels of command. It was planned to survey 10% of the Active Army SGM and their supervisors based on findings in a prior study (Dannhaus, et al., 1984).

A computer listing of all active duty SGM that contained unit addresses and selected background information was sorted by postal ZIP code. Using ZIP code clusters, installations and activities that had high densities of assigned SGM in a broad cross-section of MOSs and organizations were selected as interview sites. SGM in Reserve Component (RC) support assignments were underrepresented since many were in low density sites.

Data was collected from September 1986 through January 1987 in the continental United States (CONUS) and outside CONUS (OCONUS) in Europe, Korea, and Hawaii. Thirteen installations and activities were visited in CONUS; 18 OCONUS major commands and activities were visited. SGM and supervisors interviewed served in duty positions ranging from battalion, brigade, and post up through all levels of command to include DA activities and international commands (e.g., SHAPE HQ, Combined Field Army Korea, etc.).

RESULTS

Obtained Sample. A total of 345 SGM and 200 supervisors were interviewed. The sample included only 176 SGM-supervisor pairs. The latter was the result of a coordination problem encountered early in the data collection and later corrected.

With respect to level of organization assignment, 41.3% of the SGM and 42.2% of the supervisors were in "Division and below" assignments; 58.7% of the SGM and 57.8% of the supervisors were in assignments "Above division" (i.e., corps and higher). With respect to type organization assignment, 59.9% of the SGM and 66.3% of the supervisors were in table of organization and equipment (TOE) units; 40.1% of the SGM and 33.7% of the supervisors were in table of distribution and allowances (TDA) organizations. With respect to location, 52.6% of the SGM and 41.2% of the supervisors were interviewed at CONUS sites, and 47.4% of the SGM and 58.8% of the supervisors were interviewed at OCONUS locations.

Forty-three out of sixty-four SGM MOSs were represented in the sample. All were medium to high density MOSs. Combat arms MOSs were underrepresented since many of those SGM were in RC support assignments. However, MOS proportions obtained were representative of Active Army MOS assignments. Approximately 251 different SGM job titles were recorded. The background characteristics of the SGM sample were as follows: median age--44 years; median years military service--25 years; median time as a SGM--22 months; median time in current duty position--14 months; SMC graduates--53.4%; SGM with 1SG experience 71.4%; median number soldiers supervised as SGM--3; if SMC graduate, median time since completing SMC--32 months; average time since last "helpful" NCOES-type course--11 years; had advanced NCO course prior to attending

SMC--61.6%; and, highest levels of civilian education high school graduate (100%), "some college" (67%), "two years college" (37.3%), and "more than two years college" (16%).

Supervisor characteristics were as follows: rank - 87% field grade (major, lieutenant colonel, and colonel); most encountered supervisor rank--lieutenant colonel (46%); median number of SGM supervised in past assignments--2; median time in current assignment--10 months; median time current SGM supervised--7 months; and, percent supervised SGM SMC graduates--53.4%.

Tasks Most Frequently Performed by SGM. SGM and supervisors were asked to list tasks, duties, or responsibilities "most frequently performed" by SGM in their current job. Recorded tasks were classified into categories by a team of SGM and analysts. The five categories most frequently reported ("top five") by SGM (N=345) and by supervisors (N=199) follow. SGM "top five" categories were: (1) Office management responsibilities (41.0%); (2) Action officer on special projects (29.0%); (3) Advisor to supervisor on technical issues (28.7%); (4) Advisor to supervisor on people issues (27.5%); and, (5) Training activities administrator (24.3%). Supervisor reported "top five" tasks were: (1) Office management responsibilities (53.8%); (2) Action officer on special projects (33.2%); (3) Enlisted personnel administrator (26.6%); (4) Responsibility of soldier welfare (25.6%); and, (5) Action officer on training projects (23.6%). SGM and supervisors agree on the most frequently performed task category; however, beyond that, category names and frequencies suggest they may view the SGM's job differently.

Task Taking Most of SGM's Time. SGM and supervisors were asked to identify the one task that "took up most of the SGM's time". The "top five" task categories most frequently mentioned by SGM (N=345) were: (1) Office Management responsibilities (10.1%); (2) Inspecting and evaluating performance (4.9%); (3) Advisor to supervisor on technical issues (4.3%); (4) Unit readiness manager (4.3%); and (5) Policy and procedures administrator (4.1%). Supervisor (N=199) "top five" mentioned tasks were: (1) Office management responsibilities (17.4%); (2) Enlisted personnel manager (9.2%); (3) Overall responsibility for soldier welfare (7.2%); (4) Action officer on special projects (5.6%); and, (5) Inspecting and evaluating performance (5.1%). Again, SGM and supervisors agree on the "most time consuming task". Also note that supervisors and SGM see the SGM's job differently in this area too.

Most Important Task to Organization Mission Performed by SGM. Again, both SGM and supervisors were limited to identifying the one "most important" task performed by the SGM. The "top five" "most important task" categories reported by SGM (N=345) were: (1) Training activities administrator (7.5%); (2) Advisor to supervisor on technical issues (7.5%); (3) Office management responsibilities (7.5%); (4) Soldier welfare responsibility (7.2%); and, (5) Unit readiness manager (4.9%). The "top five" "most important task" mentioned by supervisors (N=196) were: (1) Office management responsibilities (13.8%); (2) Enlisted personnel administrator (8.7%); (3) Action officer on special projects (5.6%); (4) Advisor to supervisor on people issues (5.6%); (5) Soldier welfare responsibility (5.1%). Although, two task categories appear on both the SGM's and supervisor's "top five", SGM and supervisors differ in their perceptions of the "most important task" performed by SGM.

Most Difficult Task for SGM to Learn. Both SGM and supervisors were asked what the "most difficult task to learn" was for the newly assigned SGM. Interviewees were limited to one task. The "top five" "most difficult" tasks identified by SGM (N=344) were: (1) Learning how the organization functions (30.8%); (2) Policy and procedure administration (8.7%); (3) "Depends on SGM's background" (7.6%); (4) "None!" (4.4%); and (5) Training activities administration (4.1%). As viewed by supervisors (N=167), the "top five" "most difficult task to learn" by SGM were: (1) Learning how the organization functions (29.9%); (2) Office management procedures (8.4%); (3) "None!" (8.4%); (4) Enlisted personnel management (6.0%); and, (5) Action officer on training activities (5.4%). SGM and supervisors did agree that "Learning how the organization functions" was one of the more difficult tasks a newly assigned SGM had to learn. It is suspected that this is due to the fact that many SGM are in their first staff level assignment where tasks, duties and responsibilities may not be as well-defined as in TOE units. Furthermore, this could be a problem for an experienced SGM, because the missions, organizations, and functions of high level organizations are far from standardized.

SGM Strengths and Weaknesses Reported by Supervisors. The five most frequently reported SGM strengths by supervisors (N=190) were in the areas of: (1) "Technical ability" (33.7%); (2) "Leadership ability" (25.8%); (3) "Army experience" (25.3%); (4) "Knowing how to get things done" (21.1%); and, (5) "Concern for the soldiers' welfare" (17.4%). The five most reported "weaknesses" (N=185) were: (1) "Lack of writing ability" (27.0%); (2) "Narrow view of responsibilities" (16.8%); (3) "Tendency to be inflexible" (13.5%); (4) "Poor communicators" (10.8%); and, (5) "Too complacent" (8.6%).

What should be emphasized in the future SMC? The five subject matter areas SGM (N=343) most frequently reported should be emphasized in the SMC in the future were: (1) Army effective writing (45.1%); (2) Interpersonal communication skills (27.9%); (3) Leadership theory (21.4%); (4) How to write and present a military briefing (17.5%); and, (5) Theory of human motivation (15.6%). The five subject matter areas that should be emphasized, based on supervisor (N=184) experience were: (1) Army effective writing (41.3%); (2) How to write and present a military briefing (21.2%); (3) Staff procedures (21.2%); (4) Leadership theory (15.2%); and (5) Interpersonal communications (14.1%). In general, SGM and supervisors agree on what should be emphasized in the SMC in the future. There is a need to place emphasis on the development of communication skills, both in the areas of writing and speaking. Based on interview comments by SGM and their bosses, that is a critical SGM skill requirement. Supervisors proposed an emphasis on "Staff Procedures". That could reduce the trauma of, "Learning how the organization functions".

A Coordination Lesson Learned. It was planned to obtain a complete set of SGM-supervisor pairs. This obviously didn't happen. Initially, interviews at each site were coordinated through CSM and SGM channels. It was found that SGM were reluctant to schedule their own supervisors for interviews. A new coordination tactic was adopted. A lieutenant colonel on the team coordinated supervisor interviews through command chief of staff channels. A senior SGM coordinated SGM interviews through CSM and SGM channels. This procedure boosted SGM-supervisor pairings to nearly 100%.

DISCUSSION AND CONCLUSIONS

This study did not address what the tasks, duties, and responsibilities of SGM should be. It addressed what SGM and supervisors say SGM do on the job and what they need to know and be able to do to be effective SGM. It was assumed for the purposes of this study that SGM are doing what SGM should be doing.

The findings of the study suggest that SGM and supervisors perceive the job of SGM differently in terms of tasks observed performed, and task importance, workload, and learning difficulty. This finding is consistent with the literature of industrial and organizational psychology. This finding means that studies of this type should take experience and perceptions of both SGM and supervisors into consideration in assessing job requirements. This is essential if study findings are to be used to update or design training and professional development for senior NCOs. Both SGM and supervisors recognize what is important to do the job effectively.

The findings also showed that SGM and supervisors can agree in terms of SGM job dimensions (e.g., "Office management responsibilities", "Action officer on special projects", "Learning how this organization functions", etc.). Also, they can be in agreement in terms of training and professional development needs (e.g., "Training in communication skills, etc.").

The results suggest that the "ability to communicate in writing and to brief" don't become critical skills for most senior NCOs until they are promoted to SGM. Some SGM were criticized, because they didn't have those skills. The implication here is that development of those critical communication skills should begin earlier in the NCOES than in the SMC.

REFERENCES

- Department of the Army (DA). (1984). Individual military education and training (Army Regulation 351-1). Washington, DC: Author.
- U.S. Army Sergeants Major Academy (USASMA). (1985). Program of Instruction for United States Army Sergeants Major Course (1-250-C5). Fort Bliss, Texas: Author.
- Moreland, J. D., LaVerne, J., Neal, G. L., & Saia, F. E. (1987). Sergeant major skill and knowledge requirements (TRAC-WSMR-TEA-24-87). White Sands Missile Range, NM: U.S. Army TRADOC Analysis Command.
- Dannhaus, D. M., Neal, G. L., Roberts, P. J., Robinson, N. J., Tubbs, J. D., and Wilson, V. C. (1984). Career management field 76 training system effectiveness analysis (TRASANA TEA-19-84). White Sands Missile Range, NM: U.S. Army TRADOC System Analysis Activity.

Staff Officer Characteristics Contributing to Effective Tactical Decision Making

by

William D. Sprenger

Jon J. Fallesen

Sharon Riedel

US Army Research Institute Field Unit

Ft. Leavenworth, Kansas 66027

Eisenhower, Marshall, Pershing, Bradley...exhibited outstanding leadership qualities in leading soldiers in peacetime and combat. It was their superior decision making abilities and staff officer performance that identified them for future greatness (Puryear, 1971).

Introduction

There is considerable military anecdotal literature providing opinions, beliefs, and lore about the personal characteristics that contribute to or are required of an effective leader of soldiers. The literature does not devote as much attention to the mental and personal qualities required of a staff officer, specifically those staff officers charged with the responsibility of interpreting tactical situations, performing operational planning, making decisions, and controlling execution. To be sure, there is much overlap between the two roles - troop leader and staff officer - for an officer in the Army is expected at some time in his career to fulfill both roles. It does appear, however, that some of the characteristics that contribute to highly proficient staff performance are different from troop and combat leadership. This paper specifically examines past and on-going research of staff officer characteristics that contribute to superior tactical decision making.

Purpose

The commander and his staff make decisions at every point in the execution of their duties (even choosing not to make a decision or to delay a decision is an instance of decision making). Decision making can be improved by at least three general approaches: training, selection and aiding. Knowledge of decision maker characteristics plays a different role in each of the approaches, yet can be valuable in all. By knowing the characteristics of excellent staff officers, educational and training programs can be developed to impart those characteristics to students (to the extent that characteristics are modifiable) and to be compatible with student learning styles. Selection and placement of prospective staff officers can be sharpened based on the characteristics related to good staff performance. If we can understand how individual characteristics relate to staff performance, decision support systems can be designed to (a) overcome decision making biases, (b) fit efficient information acquisition and interpretation strategies, and (c) accommodate effective procedures to generate and select solutions. Also there is the potential to develop systems which can modify or adapt to user style characteristics (see adaptive decision aiding and intelligent interface literature, e.g. Rouse, 1988 and Halpin, 1984).

Too often behavioralists ignore individual differences or consider them some part of the uncontrolled variance in performance. This bias is contrary to several Army initiatives such as the MANPRINT program and leadership development, which advocate the examination of personnel variables that can influence performance. In any of the three approaches to improve decision making it is important to determine which individual characteristics are consistent across highly competent staff officers.

Research Approach

Review of a variety of literature sources has been made. We have examined anecdotal literature, military officer bibliographies, instructional literature of staff functions and reports of experiments which have examined the style and performance of staff officers. Based on this information we present in this paper our findings and our own speculations about the characteristics of effective staff officers.

On-going research, examining human performance in command and control by the Fort Leavenworth Field Unit of ARI, provides opportunities to collect staff characteristics data. A research method which uses a "piggy-back" approach on existing experiments makes it feasible and cost efficient to build up a data base on staff characteristics and performance.

A command and control research laboratory has been developed for the study of division level Army staff planning. Referred to as the Experimental, Development, Demonstration and Integration Center (EDDIC), the lab provides an instrumented command post to examine the use of automated information systems and decision aids by key staff positions. The procedures they used are observed and scored, as are the products, such as the development of schemes of maneuver, operations estimates, and operations orders. In addition the participants are tested or surveyed by standard psychological instruments (e.g., locus of control, cognitive abilities, and intelligence). The Officers Longitudinal Research Data Base (ORLDB) (Hunter, 1987) also allows the collection of background and experience information. Determining the diversity in staff officer profiles will be one objective of the analysis. Also characteristics will be correlated with performance on typical staff tasks to determine the strength of relationship.

The Command and Control Performance Assessment System (C2PAS) also will be pursued to evaluate staff officer performance. Using a PC-AT computer, a video disk player and related components, C2PAS presents tactical vignettes below division level in narrative and video map graphic formats. Once placed in a tactical situation a staff officer is tested in the performance of information acquisition, interpretation and tactical planning tasks. C2PAS provides a complementary test approach to EDDIC by examining the performance of an individual characteristics.

From these instrumentation opportunities and the gradual build up of data, significant relationships between staff officer characteristics and performance can be identified and new hypotheses generated.

Findings

Thompson described his views on the characteristics of a "high performing staff" (1984a) and what it takes to develop and sustain one (1984b). He believed that competent staffs display the following characteristics: goal clarity, teamwork, focused energy, knowledge and procedures, creative standardization (routine procedures that do not hamper creativity or effectiveness), meta-language (a shared, staff-peculiar efficient language), innovation, rehearsal, rhythm (smooth progression of events), core values (e.g., producing a quality product, ethics, professionalism), reputation and adaptability.

Any good team is comprised of competent individuals. The elements of staff competency have been addressed by various authors. Newman, who spent a good portion of his military career as a staff officer, largely because he was recognized for his superior performance, described the essential characteristics of a staff officer as one who displayed knowledge, infinite tact, delicacy of judgement and flexibility (1981). He attributed the same characteristics to good commanders. Johnson and Walker (1975) summed up staff officer success as preparation, application and attitude. They identified the following performance skills as being essential for quality staff officer performance: good written and oral communication, listening ability, concentration, high reading ability with comprehension, good observation skills, knowledge of subject matter, excellent writing skills, organization, large and precise vocabulary, formal and informal speaking ability, memory, and ability to follow through on tasks. Thompson (1984a) contended that individuals from "high performing staffs" have these characteristics: core values of staff, acceptance into the group, task enjoyment, routine or unconscious performance of tasks, obsession with job, social closeness, and measurement of time by activities.

Thompson (1984b) also described four human dimensions which can be used to determine the level of proficiency of a staff: adaptability, goal setting, job maturity, and psychological maturity. He differentiated between job and psychological maturity; the former being the ability to perform a particular task (e.g., problem solving doctrinal knowledge, task relevant skills, communication ability, knowledge of staff procedures) and the latter as the willingness, confidence and commitment to perform.

Michel and Riedel (in press) investigated the effects of level of expertise and cognitive style on the development of a concept of operations. There was no significant relationship between cognitive style, as measured by the embedded figures test, and the usage of tactical information. Although the concept of operations which were developed did not vary substantially, the staff officers who were considered experts solved the problem using more summary and less detailed information than the novices did. Both the lack of a cognitive style effect and the existence of an expertise effect were consistent with previous studies.

Wojdakowski (1988) presented a list of desirable staff characteristics as a model based on coordination, consistency and credibility. He believes that staff success requires efficient completion of staff functions which in turn rely upon good time management and "knowing" when to intervene. Accuracy is key also. Accuracy in the uncertain environment of war comes from precision in doing little things well. To ensure accuracy of staff efforts, Patton would have his staff officer go to the front at least once a week to obtain specific information so they could determine first hand feedback about their past successes and mistakes. Typically, accuracy requires time. If the staff is going to be efficient they have to adapt their procedures to attain sufficient levels of accuracy. Flexibility is needed to detect and correct errors. To help ensure proper use of time, Wojdakowski advocates consistent means of analysis and reporting. Familiarity of the process and format will aid both the conduct of the task and communication of information to others.

Lussier, Solick, and Garlinger (1987) found in a evaluation of staff group problem solving that poor performance could be attributed to poor estimates, insufficient analysis, ignoring critical analysis criteria, and failure to check for errors. Also they found that poorer performing groups did not use rough estimation techniques as they should have to scope the problem or to prune unpromising branches. Queries for information were addressed either (a) quickly in too vague a manner and then dropped or (b) slowly in excessive detail. In an analysis of several command episodes through history, Van Creveld (1935) pointed out that success was determined largely when command activities sought focused information to augment subordinate's.

Our Construct

Through the review of literature like that described above, over 100 individual difference variables for the study of staff officer performance were compiled. After digesting the literature and data some of the key variables were identified to examine in initial research. The goals of the research will be to determine the extent of the relationships of hypothesized staff characteristics on staff tasks, such as generation of alternative courses of action, evaluation of courses of action, recommended course of action, production of an operations estimate, development of plans and orders, and control of battle execution.

The characteristics that we hypothesize to lead to staff officer competency are:

Adaptive--

- a. has fluid intelligence, adaptive to task requirements;
- b. has the ability to synthesize information and generalize to future situations.

Experienced, Intelligent--

- a. has crystallized intelligence, maintains an in-depth knowledge of functional area responsible for, experience in working at subordinate levels, knowledgeable about the nature of related staff areas and the relationships among various staff functions;
- b. has extensive experience in task procedures and analytical techniques, knows when it is appropriate to apply them;

Analytical--

- a. relies on evidence-based decisions;
- b. applies analysis in level of detail appropriate for the importance of the problem and the time available, uses approximation techniques as appropriate, matches level of precision to the criticality of subproblem and the time available;
- c. is not solely reactive to commander but anticipates commanders needs, extension of commander, is aware of biases and shortfalls, counteracts them as appropriate;

Assertive--

- a. holds a "healthy" skepticism, avoids "groupthink";
- b. second guesses information and products of self and others, checks on accuracy;
- c. does not persevere on proving correctness of own beliefs or positions, accepts ideas and opinions of others;

Interpersonal Skills--

- a. adaptive to commander's leadership style;
- b. communicates well as demonstrated by listening and presentations;
- c. flexible in dealing with others.

Final Remarks

The literature is sprinkled with characteristics of staff officers; most of it limited to supposition (just as what we have offered above). Some of it has been confirmed through historical review, some by first-hand observation and some even by structured evaluation and quantification techniques. Through our efforts we hope to determine if it is possible to distinguish among characteristics for effective staff performance and to develop recommendations for improving training and enhancing decision making procedures based on these individual characteristics.

References

- Halpin, S. M. (1984). A proposal for an intelligent interface in man-machine systems. Proceedings of the 23rd IEEE Conference on Decision and Control, 592-595.
- Harte, J. E. (1987). Officer longitudinal research data base: Development and utilization. Proceedings 29th Annual Conference of the Military Testing Association, 90-95.
- Lussion, J. W., Solick, R. E., & Garlinger, D. K. (1987). Measurement of group planning and problem solving abilities. Proceedings 20th Annual Conference of the Military Testing Association, 363-367.
- Michel, R. R. & Kiedel, S. L. (in press). Effects of expertise and cognitive style on information use in tactical decision making. Army Research Institute Technical Report 806. Alexandria, VA: ARI.

- Newman, A. S. (1981). Follow Me, The Human Element in Leadership. Novato, CA: Presidio Press.
- Puryear, E. F., Jr. (1971). 19 Stars, A Study in Military Character and Leadership. Novato, CA: Presidio Press.
- Rouse, W. B. (1988). Adaptive aiding for human/computer control. Human Factors, 30(4), 431-444.
- Thompson, H. L. (1984a). High performing staff part I: What is it? Army Organizational Effectiveness Journal. 8(1), 1-15.
- Thompson, H. L. (1984b). High performing staff part II: Developing and sustaining the HPS. Army Organizational Effectiveness Journal.
- Wojdakowski, W. (1988). A staff philosophy. Military Review, LXVIII(11), 43-52.
- Van Creveld, M. (1985). Command in War. Cambridge: Harvard University Press.

Vietnam: Lasting Effects on Confidence Toward Military Leaders

Frank J. Ricotta, Jr. & Charles N. Weaver
St. Mary's University

Michael D. Matthews
Drury College

"Hell no we won't go" and "Ho Ho Ho Chi Minh, the NLF is going to win" are two extreme examples of attitudes adopted by young American citizens during the controversial Vietnam conflict. Those who were draft age during that era were particularly vocal in their opposition to the war. Coupled with the attention that the media gave to such views, and the lack of support for the war among the general population, one might question the extent to which the attitudes of those persons most affected by the war—draft age Americans—were and possibly continue to be affected by it.

Confidence toward the military and its leaders is one such attitude that may have been negatively affected by the Vietnam war. Attitudes toward our nation's leaders and the military have been assessed regularly (Department of Commerce, 1977, 1980; Ladd, 1977), and show that such attitudes are influenced by national events in general, and the Vietnam war in particular. However, these reports have not analyzed the attitudes of any specific age cohorts. The current paper compares attitudes toward the military of persons—both male and female—who were draft age during the Vietnam war with a comparison age group of older Americans that were not draft age during that period. Furthermore, the attitudes of these two groups were compared yearly from 1973 to 1984, with the exception of 1979 and 1981 when such data were not collected. It was hypothesized that draft age persons would show less confidence in the military than the comparison group, but a prediction of how much the attitudes of the two groups may have changed over the years since the end of the war is not clear. The current study employs a cross-sectional design, and therefore cannot separate out cohort effects from period effects. Determining which of these most influenced attitudes would require a longitudinal design (Gleim, 1976).

Method

The data used for this study come from the General Social Surveys conducted by the National Opinion Research Center. Data from the 1973, 1974, 1975, 1976, 1977, 1978, 1980, 1982, 1983 and 1984 surveys were included for analysis in the current study. Each survey consists of an independently drawn sample of English-speaking persons 18 years of age or older, living in non-institutional arrangements within the continental United States (Davis, 1984). The total N for the draft age cohort (subjects who were age 18-25 in 1973) was 2761. For the comparison group, who were age 34 to 41 in 1973, the N was 1790.

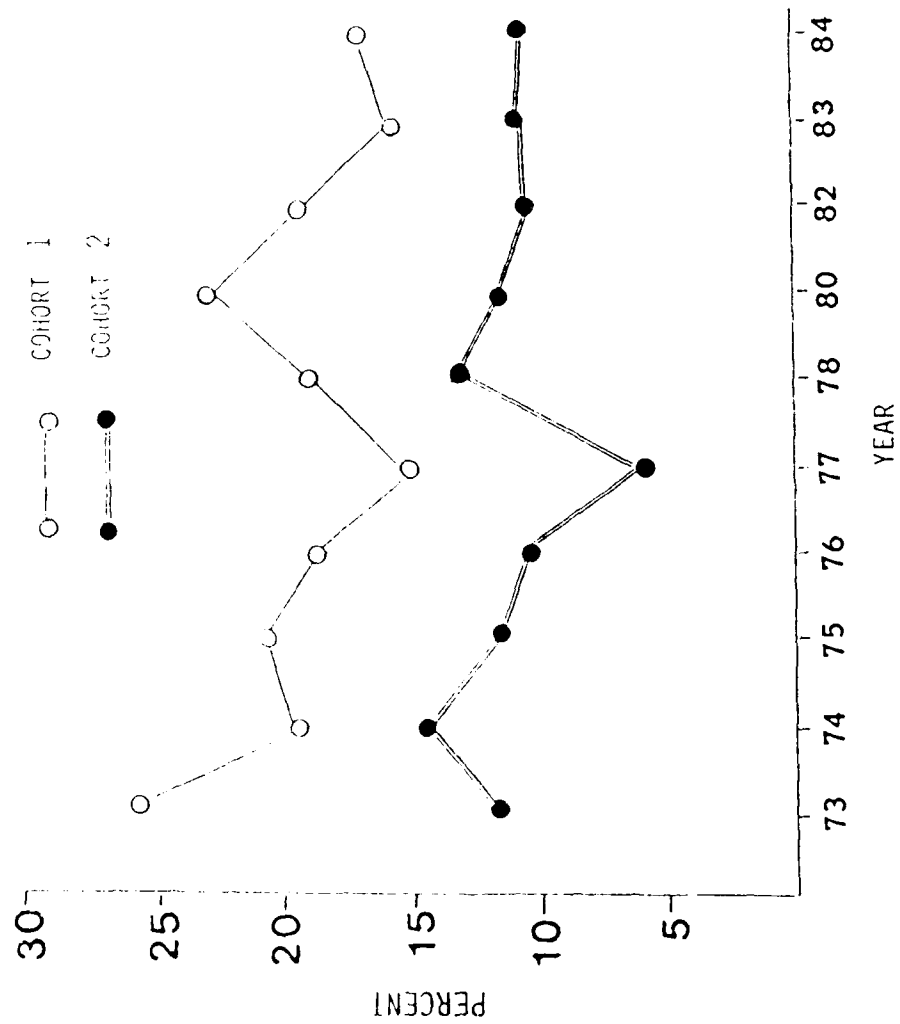
The respondents were asked to indicate how much confidence they had in people running various institutions, including the military. They were required to indicate whether they had "a great deal," "only some," or "hardly any" confidence toward military leaders. Cross-tabulations relating cohort group to confidence in military leaders were examined using a format and analysis plan derived from Glenn and Zody (1970).

Results and Discussion

In order to examine the hypothesis that the draft age cohort (cohort 1) would display less confidence toward military leaders than the comparison group (cohort 2), the percentage of each cohort indicating that they had "hardly any" confidence in military leaders for each of the years for which data were collected was compared. Tables showing percentage of respondents indicating "a great deal" or "only some" confidence in military leaders are not included in this paper due to space limitations, but are available from the third author upon request.

Figure 1 shows this relationship. Examination of this figure clearly indicates that, compared to cohort 2, draft age persons had substantially less confidence in the military in 1973, near the end of the war, and that their lack of confidence in the military has persisted over the years. Although there have been fluctuations in confidence toward military leaders for both cohorts, these changes tend to covary with the absolute difference between the two groups remaining fairly constant. The greatest difference in magnitude between the two cohorts on this measure was 13.8 percentage points in 1973, and the least was 4.3 in 1984. These differences for each of the years studied, beginning with 1973 and ending with 1984 were 13.8, 5.5, 8.8, 7.4, 8.6, 5.1, 10.4, 7.2, 4.3, and 6.3.

It is interesting to speculate what events may have influenced fluctuations in confidence toward military leaders. The least confidence, shown by both groups in 1973, can reasonably be attributable to the Vietnam war itself. Confidence among members of both groups then improved steadily, reaching a minimum level of disapproval in the middle 1970s, before a sharp trend toward greater disapproval, which peaked again in the early 1980s. These fluctuations are of substantial magnitude. The mean percentage of subjects collapsed across age cohorts, indicating "hardly any" confidence toward the military was 16.5, 10.6, and 16.9 for the years of 1973, 1977, and 1980, respectively. Interestingly, each of these years which showed extremes in attitudes were either an election year or the year following an election year. Perhaps campaign rhetoric heightens public awareness of national defense issues, and this heightened awareness interacts with contemporary events to influence attitudes toward the military. A full discussion of this topic is beyond the scope of this paper, but perhaps the data presented here will serve a heuristic function and stimulate further thinking about causes and/or correlates of these trends.



The persisting nature of the attitudes of the draft age cohort also raises some interesting areas of speculation. Does their lesser degree of confidence in military leaders generalize to less support for the defense establishment as a whole? Might they be more critical of attempts by the nation's leaders to involve its armed forces in future regional conflicts? Are they less supportive of increases in the defense budget for new weapons systems or enhancement of conventional forces? Do they tend to prefer a foreign policy which avoids direct manipulation of the internal affairs of other countries?

The responses to such empirical questions carry a great deal of importance because people who were draft age during the Vietnam war are now approaching the age at which they will have the greatest influence in the major institutions of the nation. As they become leaders of business, education, religion and the military, their attitudes will have a greater chance of being manifested overtly in public policy, such as the draft. Further investigation of the impact of the Vietnam war on social and political attitudes and behavior appears warranted.

References

- Davis, J.A. (1984). General social surveys, 1972-1984: Cumulative codebook. Chicago: National Opinion Research Center
- Glenn, N.D. (1976). Cohort analysts' futile quest: Statistical attempts to separate age, period, and cohort effects. American Sociological Review, 41, 900-904.
- Glenn, N.D., & Zody, R.E. (1970). Cohort analysis with national survey data. The Gerontologist, 10, 233-240.
- Ladd, E. C., Jr. (1977). The polls: The question of confidence. Public Opinion Quarterly, 40, 545-552.
- U.S. Department of Commerce. (1976). Social indicators, 1976. Washington, D.C.: Government Printing Office.
- U.S. Department of Commerce. (1980). Social indicators III: Selected data on social conditions and trends in the U.S. Washington, D.C.: Government Printing Office.

PREDICTING LEADERSHIP AT THE SERVICE ACADEMIES AND BEYOND

Dr. Robert F. Priest, Chair
U.S. Military Academy
West Point, New York

This panel reported on recent research, at the U.S. service academies, on the usefulness of biographical data in predicting leadership at the academies or after commissioning. The first paper outlines the development and experimental use of a biographical inventory at the Naval Academy to assess several to assess several personality traits that are frequently associated with leadership. The next paper describes the use of structured interviews and life history essays to generate biographical items relevant to leadership at the Naval Academy, and the validity of the resulting questionnaire in predicting military performance at the academy. The usefulness of human development theory in guiding the construction of biographical measures is also discussed. The third paper reports on a pilot study of Coast Guard Academy graduates that examined the association between biographical data obtained at the time of application to the Academy and performance two years after commissioning. Dr. Robert F. Priest discussed informally the validity of biographical data, particularly high school teachers' ratings, in predicting leadership at the Academy and promotions after commissioning. This research was reported at the 1987 Annual Conference of the Military Testing Association. Dr. Martin F. Wiskoff provides comments on the presentations and their policy implications. A general discussion ensued at the panel that emphasized the importance of clarifying, both conceptually and operationally, the meaning of leadership.

Assessing Leadership Potential at the Naval Academy with a Biographical Measure

Lawrence J. Stricker

Educational Testing Service

Applicants to the Naval Academy are intensively screened. In recent years, the focus of the screening was on identifying applicants who would be successful students. However, in 1984, the Secretary of the Navy directed that the emphasis be shifted to identifying applicants who would be future leaders. This study represents one effort in this direction: the development of a biographical measure to assess the leadership potential of Academy applicants.

A biographical measure was chosen for this purpose because such devices have distinct advantages over personality and interest inventories, and similar instruments. Biographical measures capture directly the past behavior of a person, probably the best predictor of his or her future actions. And the measures deal with facts about the person's life, not the introspections and subjective judgments that make up the content of personality inventories and the like. As a result, biographical measures are likely to be less prone to misinterpretation, resistance, and distortion.

Recent reassessments of the empirical research on the personality correlates of leadership ability suggest that this ability can be successfully predicted: consistent links appear to exist between personality traits and leadership (Hogan, 1987). This reassessment calls into question the widely-held conclusion that personality traits and other individual-difference variables do not distinguish leaders from followers, and that leadership is simply a function of the situation (Stogdill, 1948; Gibb, 1954). This reappraisal of the work on leadership also raises the real possibility that a properly constructed biographical inventory, designed to measure relevant personality traits, may be able to assess leadership potential.

Accordingly, the purpose of this study was to construct and validate a biographical inventory to measure personality traits that are predictive of leadership.

¹This research was supported in part by the U.S. Navy-American Society for Engineering Education Summer Faculty Research Program (assembling the biographical inventory), and in part by the Navy Personnel Research and Development Center, U.S. Army Research Office, and Battelle Institute under contract DAAL03-86-D-0001 (item and validity analyses of the inventory). The views, opinions, and/or findings contained in this report are those of the author and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other documentation.

Method

Constructing the Inventory

Content. The empirical research on the personality correlates of leadership was reviewed. Because this literature is massive and has been extensively reviewed, the reviews themselves were reviewed. Five personality traits, dominance, emotional stability, need for achievement, self-confidence, and sociability, were identified as being more-or-less consistently related to leadership in the reviews.

Items were written to tap each of the five traits. The items have these characteristics:

1. The items use multiple-choice or Yes-No formats.
2. The alternatives for multiple-choice items are on continuous scales. This feature facilitates quantification of the responses.
3. The items (stems and alternatives) are factual. This characteristic accords with the basic conception of a biographical item.
4. The items deal with public behavior. The factual nature of the items, in combination with the public behavior inquired about, makes the responses, in principle, verifiable; the verifiability is expected to reduce distortion.
5. The items (a) concern behavior that is under the examinees' control and (b) involve opportunities and resources available to virtually everyone. These characteristics are expected to enhance the validity of the items, while minimizing unfairness and bias.
6. Apart from a few items dealing with easily recalled activities that inquire about whether these things occurred at any time in the examinees' lives, most items concern a particular time period: during high school, during the senior year, or since entering the senior year. This practice standardizes the period being described, and focuses on a recent period that can be recalled accurately and is likely to be most relevant to the examinees' current behavior.

Item analysis. An inventory made up of the five tentative personality scales and a Social Desirability scale was administered on the second day of 1987 Plebe Summer to the 642 entering midshipmen in one of the two battalions of plebes. (Plebe Summer is a seven-week bootcamp for the training and indoctrination of incoming midshipmen.) The midshipmen were instructed that the results would not become part of their official records.

The item analysis of the inventory was carried out for one random half (N=233) of 1987 high school graduates--the other half of the 1987 high school graduates (N=233) and the earlier graduates (N=171) were reserved for the analysis of the inventory's validity.

The item analytic procedures (Jackson, 1970) were intended to (a) maximize convergent and discriminant validity; and (b) ensure that response styles, and sex and ethnic-group bias, were minimized.

The item analysis focused on each item as a whole, not on individual response alternatives. Alternatives for multiple-choice items were dichotomized at the median. Items were scored 0 or 1, with all responses in the same dichotomy being assigned the same score.

Items were eliminated if (a) they had extreme endorsement frequencies (less than .05 or greater than .95); (b) they did not correlate significantly ($p > .05$) with their own scales; or (c) they had higher correlations with one of the other personality scales, the Social Desirability scale, an ad hoc acquiescence measure, ethnicity, or sex.

The number of items analyzed ranged from 42 to 60 for the five personality scales; the number of items on the final scales ranged from 8 to 22.

Validity Analysis

The validity analysis was intended to assess the ability of the final forms of the personality scales to tap leadership as well as the personality traits that they were intended to measure. (It is conceivable that a scale may validly measure the intended trait but be unrelated to leadership, at least as it is manifested in the setting being studied.) This analysis was also designed to appraise the involvement of response styles in the personality scales, and sex and ethnic-group differences on the scales.

As noted earlier, the analysis was done separately for the random half ($N = 233$) of 1987 graduates not used in the item analysis and for earlier graduates ($N = 171$). Only the results for the 1987 graduates will be discussed because the inventory was not designed for use with individuals who are not recent graduates.

Criteria. Peer ratings were the criteria for the personality traits, and the primary criterion for leadership. Secondary criteria of leadership were two Academy grades, Military Performance and Professional Military Quality Point Rating, which may reflect leadership as well as other variables.

The peer ratings were obtained on the last weekend of Plebe Summer for the midshipmen in the battalion that had been given the inventory. The midshipmen were asked to rate themselves and the other members of their squad on the five personality traits plus leadership. The poles of the variables were defined, and an eight-point scale was used. The ratings were made anonymously, and the midshipmen were assured that the ratings would be used for research purposes only. The median rating received by each midshipman was standardized within squads.

The Military Performance grade was assessed at the end of the first semester by the midshipman's company officer, based on performance in a variety of areas, such as drill and parades, sports participation, personal appearance and military bearing, and academic performance.

The Professional Military Quality Point Rating was an average of first-semester grades for Military Performance, Conduct, Physical Education, and professional courses.

Other variables. Other variables in the analysis were sex, ethnicity, and Acquiescence and Social Desirability scales made up of items similar in form to those on the personality scales.

Results and Discussion

Intercorrelations of Personality Scales, Response Style Scales, Ethnicity, and Sex

The intercorrelations of the personality scales, response style scales, sex, and ethnicity are reported in Table 1. All the scales correlated positively or near zero with each other, except for Self-Confidence, which correlated negatively with Need for Achievement Dominance and Sociability correlated substantially with each other, and Need for Achievement correlated moderately with Dominance and Sociability.

The Emotional Stability and Self-Confidence scales correlated moderately with the Acquiescence scale. And the Emotional Stability, Need for Achievement, and Sociability scales correlated moderately with the Social Desirability scale. These correlations imply some response style involvement in the personality scales, despite the precautions taken in the item analysis.

All the scales correlated near zero with sex and ethnicity, suggesting that bias is absent from them.

Correlations of Personality Scales with Criteria

The correlations of the personality scales with the criteria are reported in Table 2. The Dominance and Sociability scales correlated moderately with their corresponding peer ratings, and the Need for Achievement and Self-Confidence scales correlated modestly with the corresponding ratings. However, the Self-Confidence scale correlated higher with an irrelevant rating than with the corresponding rating.

The Sociability scale correlated moderately with the Leadership rating. And the Sociability and Need for Achievement scales correlated modestly with Military Performance and Professional Military Quality Point Rating.

In short, there was some evidence of convergent and discriminant validity for three scales (Dominance, Need for Achievement, and Sociability). Furthermore, one scale (Sociability) had some validity in predicting leadership.

Table 1
Intercorrelations of Personality and Response

Style Scales, Sex, and Ethnicity

Variable	1	2	3	4	5	6	7	8	9
1. Dominance scale	(.68)	-.12	.39	-.09	.57	.19	.22	-.02	-.04
2. Emotional Stability scale		(.68)	.04	.23	-.06	-.25	.28	-.01	-.08
3. Need for Achievement scale			(.66)	-.22	.44	.09	.36	-.06	-.10
4. Self-Confidence scale				(.67)	-.16	-.34	.14	-.06	.15
5. Sociability scale					(.78)	.10	.37	.04	-.03
6. Acquiescence scale						(.58)	-.27	-.02	.05
7. Social Desirability scale							(.27)	-.14	-.11
8. Sex ^a								(--)	.03
9. Ethnicity ^b									(--)

Note. Coefficient Alpha reliability coefficients appear in parentheses. N is 233. Correlations of .13 and .17 are significant at the .05 and .01 levels (two-tail), respectively.

^aMale=1, Female=0 ^bWhite=1, All others=0

Table 2
Correlations of Personality Scales with Criteria

Personality Scale	Rating							
	Dominance	Emotional Stability	Need for Achievement	Self-Confidence	Sociability	Leadership	Military Performance	HQPR ^b
Dominance	.23	.01	.06	.13	.20	.06	.01	.04
Emotional Stability	-.11	.03	.04	-.09	-.19	.06	-.02	.02
Need for Achievement	.08	-.01	.12	.03	.07	.07	.12	.13
Self-Confidence	.10	.14	.15	.12	.03	.09	.03	.07
Sociability	.35	.13	.21	.27	.36	.28	.13	.12
Interrater Reliability ^a	.87	.76	.86	.86	.88	.89	--	--

Note. Ns vary from 199 to 223. For an N of 223, correlations of .13 and .17 are significant at the .05 and .01 levels (two-tail), respectively.

^aEstimated by the correlation between the ratings for random halves of the raters, corrected for double length by the Spearman-Brown formula. ^bProfessional Military Quality Point Rating

Conclusions

A key finding is that the Sociability scale, and to a lesser extent, the Dominance and Need for Achievement scales, had some validity. However, the level of validity was often modest. The failure of the Emotional Stability and Self-Confidence scales to show any sign of validity cannot be explained at this juncture.

The .28 correlation between the Sociability scale and the Leadership rating for 1987 graduates, coupled with its small correlations with the secondary criteria of leadership, suggests that this scale may be useful in selection, at least for new high school graduates. The value of the scale for this purpose clearly requires further confirmation. The scale needs to be administered under operational conditions to applicants and its validity appraised in that context, including its incremental validity vis-a-vis current selection measures.

In interpreting the validity results, it must be borne in mind that the ratings and the leadership criteria were less than ideal. Other analyses indicate that the ratings, including the primary leadership criterion, were affected by a halo factor. And the secondary leadership criteria, Military Performance and Professional Military Quality Point Rating, reflect things besides leadership.

The present results, in total, offer no more than modest support for the proposition that personality traits are implicated in leadership. However, this conclusion must be qualified because of the methodological limitations already noted and the specialized nature of the leadership situation being studied: leadership by incoming midshipmen. Whether the present conclusions are generalizable to other contexts in the Academy, in the Navy, or elsewhere is uncertain. As a first step, follow-up studies of the predictability of leadership in other situations at the Academy would be valuable.

References

- Gibb, C. A. (1954). Leadership. In G. Lindzey (Ed.), Handbook of social psychology. (Vol. 2, pp. 877-920). Cambridge, MA: Addison-Wesley.
- Hogan, R. (1987). The return of the repressed: A theory of social action. [Review of Personality in the social process.] Contemporary Psychology, 32, 43-44.
- Jackson, D. N. (1970). A sequential system for personality scale construction. In C. D. Spielberger (ed.), Current topics in clinical and community psychology. (Vol. 2, pp. 61-96). New York: Academic Press.
- Stogdill, R. M. (1948). Personal factors associated with leadership: A survey of the literature. Journal of Psychology, 25, 35-71.

Predicting Leadership at the
Service Academies and Beyond

Craig J. Russell
Institute of Management and Labor Relations
Rutgers, The State University
New Brunswick, NJ 08903
(201) 932-9022

Karl W. Kuhnert
Department of Psychology
University of Georgia
Athens, GA 30602
(404) 542-8891

The major goals of the U.S. service academies involve the identification and development of future military leaders. These goals are achieved through careful student selection procedures followed by a rigorous curriculum of scholastic and military education. Selection procedures usually involve some combination of prior scholastic achievements, measures of vocational preference, and non-scholastic activities. Developmental activities include both the academic curriculum and military training components of academy life.

Hunter and Hunter's (1984) recent meta-analysis of personnel selection instruments found that selection technologies based on the consistency principle are best at predicting subsequent job performance. The consistency principle states that past behavior and performance is the best predictor of future behavior and performance (Wernimont & Campbell, 1968). The key question for service academies is: What kind of high school and service academy experiences predict subsequent performance as leaders in the military?

Recent work in the area of biographical information in personnel selection holds implications for both student selection and development. The purpose of this paper is to 1) describe alternate methods for identifying critical life events that predict subsequent performance as military officers and 2) briefly outline implications for developmental activities at the academies.

Biographical Information and Military Leadership
Theoretical Rationale

Biographical information, or biodata, has been repeatedly found to be the most powerful and lowest cost predictor of job performance in a wide variety of settings (Owens, 1976). Owens (1968, 1971) developed a Developmental-Integrative (D-I) model to explain why biographical data predicts subsequent performance. Specifically, Owens argued that prior life experiences represent critical points in an individual's development and, by systematically capturing aspects of these situations, biodata can identify those individuals who have had similar experience patterns. The nature and content of these prior developmental episodes cause the people exposed to them to learn, grow, and develop in similar ways. It is not the exposure to a life

- and objectives (career and personal);
2. have insight into the goals and objectives of his crew and subordinate officers, his peers in the fleet, and his superior officers;
3. know what constitute appropriate and inappropriate ways of influencing the crew, subordinate officers, peers, and superiors in the process of pursuing the various goals and objectives (the rules of exchange); and,
4. have a thorough understanding and belief in the overarching values and laws that govern our country.

Further, the fully developed leader would be expected to know when the values and laws that govern our country supercede his personal goals and any commitments he has made to others (e.g., while a captain is responsible for the well-being of his crew, he/she has to know when it is appropriate to risk that well-being for the good of the country).

Procedures

While it is important to know a leader's stage of development, it is also critical to understand how and in what ways past life experiences influence that development. Three recent studies have demonstrated ways of capturing life experiences that relate to performance in leadership positions.

Russell, Mattson, Devlin, and Atwater (1988) reported the results of a study demonstrating that systematically captured experiences from high school predict leadership performance ratings at the U.S. Naval Academy. Specifically, they had 4th class midshipmen (freshmen) complete life history essays describing individual and group accomplishments, disappointing situations, and stressful situations they had encountered in high school. Each question was accompanied by follow up questions targeted at various dimensions of the officer fitness report forms (e.g., resource management, working relations, etc.). Written responses by 70 midshipmen were content analyzed and developed into 100 biographical questions (e.g., How often have you failed to achieve some goal because you were initially overconfident?). Responses to these questions by the class of 1991 predicted subsequent military performance ratings at the Naval Academy.

Russell (1988) and Lindsey, Homes, and McCall (1987) independently developed structured interview procedures for predicting performance of top level corporate leaders. Lindsey et al. interviewed 86 successful top level executives from 6 firms (5 in the Fortune 50) regarding key events in their lives and what they learned from them. They were able to identify a number of categories of life experiences (e.g., developmental assignments, hardships, etc.) and lessons learned. Unfortunately, since these were all "successful" executives, we can't be sure that the unsuccessful ones did not have the same experiences and learn the same lessons.

However, Russell (1988) reported a similar interview procedure conducted on 66 candidates for top level management positions in a Fortune 50 firm. Ratings made from the interviews predicted measures of the candidate performance. The interview procedures described by Russell (1988) and Lindsey et al. (1987) both involve gathering examples of prior life experiences, what the subject felt

at the time (cognitions, affect, attitudes, etc.), what they did (behaviors), outcomes of the situation, and environmental characteristics. Both procedures involved the use of categories to classify the components of prior experience. Russell's (1988) procedure additionally demonstrates that ratings can be made that predict performance outcomes.

Future efforts need to explore different:

1. types of stimulus materials (e.g., types of interview questions, life history questions);
2. ways of asking for biographical information (e.g., paper and pencil test vs. interviews with Naval Academy Information Officers); and,
3. type of biographical information (e.g., how a life experience might have different meaning depending on the stage of a person's development).

Developing Leaders

Biographical information has implications for both the identification and development of military leaders. The identification of "critical midshipmen events" that predict subsequent performance as officers may have numerous implications for subsequent guidance of cadets and midshipmen activities at the military academies. For example, one of the first leadership opportunities for a midshipmen is in their interaction with plebes. Knowing which upperclassmen perceptions, feelings, and behaviors in their interactions with plebes are related to subsequent performance as an officer will help guide faculty and staff to maximize the developmental opportunities for upperclassmen.

Further, one would expect that midshipmen or cadets at different developmental stages are characterized by different profiles of leadership skills and have predictably different ideas about how to interact with subordinates, peers, and superiors. For example, one experience may involve having to deal with situations where people in authority expect conflicting things from you (one of the biodata questions developed by Russell et al., 1988 reflecting high school experiences). If this type of experience and what was learned from the experience predicts performance as a junior officer, it implies that:

- 1) midshipmen should be systematically exposed to situations like this;
- 2) faculty and staff at the military academies should be trained in how to manage these situations to ensure that the midshipmen and cadets get the maximum amount of development out of the experience (both in terms of managing the situation and in providing feedback and counseling to the student); and,
- 3) evaluation systems at the military academies should be modified to reflect observations of these experiences.

Midshipmen and cadets would be exposed to those situations most appropriate for their stage of development.

Knowing what experiences are most important and what a cadet or midshipman needs to be getting out of a particular experience would enable the military academies to ensure that students 1) receive the kind of challenge/training most appropriate for the development of military leaders and 2) get the most out of each

opportunity or challenge presented them.

References

- Avolio, B.J., Waldman, D.A., & Einstein, W.O. (1988) Transformational leadership in a management game simulation. Group and Organizational Studies, 13, 59-80.
- Bass, B.F. (1985) Leadership and performance beyond expectations. New York: Free Press.
- Burns, J.M. (1978) Leadership. New York: Harper Torchbooks.
- Hunter, J. & Hunter, R. (1984) Validity and utility of alternative predictors of job performance. Psychological Bulletin, 96, 72-98.
- Kegan, R. (1982) The evolving self: Problems and process in human development. Cambridge, Mass.; Harvard University Press.
- Kuhnert, K.W. & Lewis, P. (1987) Transactional and transformational leadership: A constructive/developmental analysis. Academy of Management Review, 12, 648-657.
- Lindsey E., Homes, V. & McCall, M. (1987) Key events in executive lives. Technical report #32, Center for Creative Leadership: Greensboro, North Carolina.
- Owens, W.A. (1968) Toward one discipline of scientific psychology. American Psychologist, 23, 782-785.
- Owens, W.A. 1971 A quasi-actuarial prospect for individual assessment. American Psychologist, 26, 992-999.
- Owens, W.A. (1976) Background data. In M.D. Dunnette (Ed.), Handbook of industrial and organizational psychology: 609-644. Chicago: Rand McNally.
- Russell, C.J. (1988, August) Biographical information generated from structured interviews for the selection of top level managers. Presented at the annual meeting of the Academy of Management, Anaheim, CA.
- Russell, C.J., Mattson, J., Devlin, S.E., & Atwater, D. (1988) Predictive validity of biodata items generated from retrospective life experience essays. Presented at the 3rd annual Society for Industrial and Organizational Psychology meetings, Dallas, TX.
- Wernimont, P. & Campbell, J. (1968) Sign, samples, and criteria. Journal of Applied Psychology, 52, 372-376.

The U. S. Coast Guard Academy
Class of 1986:
Biographical Predictors of Success

Earl H. Potter and Robert R. Albright
United States Coast Guard Academy

The Coast Guard is a small service in comparison with the Army, Navy and Air Force. The Coast Guard Academy has a cadet corps in the neighborhood of 850 down from a peak of 1200 in 1976. Perhaps because we are small, we tend to look at biographies rather than biographical data. Our use of biographical data can best be described as intuitive. Each year the Superintendent issues a precept to the Cadet Candidate Evaluation Board (CCEB) containing his guidance for the selection process. In the Fall of 1981 the Director of Admissions had prepared a summary of his advice to the Superintendent which included the following statement about the value of after school work as a predictor of success at the Academy: "...the candidate who has a record of strong work habits or useful application of his or her time should be successful at the Academy." In 1988 the direction to the CCEB included the same advice despite the fact that some research, for example, Willingham's (1985) work for the Educational Testing Service has shown that work experience is not predictive of success in college.

Part of the reason for a focus on surface validity is the Academy's fundamental need for a specific array of talents in the corps. We "need" to field a football team, soccer team, volleyball team and 16 other intercollegiate teams. We need flute players for the band, talkers for the debating team and editors for the yearbook. Much of our interest in biographical data represents a concern for balance in the corps. Participation in multiple activities may predict success in college; it is an essential attribute at the Coast Guard Academy. Any selection system favoring the best predictors of success over the discretion to take a risk on a candidate with a desirable skill would not be acceptable to the Coast Guard.

Another reason for our approach has to do with a problem which we share with any other organization that wants to identify leadership potential--the problem of a valid success criterion. David Campbell co-author of the Strong-Campbell Vocational Interest Inventory reports that he discovered several years ago that the Naval Academy was using his test as a selection criterion. The Navy had identified the profile of those graduates more likely to remain on

active duty following the expiration of the officer's initial obligation and was including a factor that favored that profile in their selection criteria. Campbell worried that this profile which emphasized "military" values was appropriate for mid-grade officers but might not fit the future Admirals of America's 21st Century Navy. At the Coast Guard Academy we lose a number of good people each year because they experience a "change in career objectives." To translate--they don't like the way we do business and don't see opportunities ahead that persuade them that endurance is worth the free education. We may decide in the future that we do some things at the Academy that drive out the very people we want to keep. Graduation from the Coast Guard Academy may not be the criterion upon which you would want to build a selection device that would shape the future of the Coast Guard.

C. Paul Sparks who for 30 years directed research on management potential at the Exxon companies has found (Sparks, 1988) correlations of .63 and .50 (in samples of 222 and 221, $p < .01$) between a background questionnaire and success in management. In a follow-up study 12 years later the correlation between background and success was still .34 ($n=304, p < .01$). Terry W. Mitchell working first at LIMRA and now at the Psychological Corporation has developed sophisticated on-line and paper and pencil biographical instruments that rule out the possibility of misrepresentation of biographical data almost entirely. We are aware that it is possible to design biodata instruments that improve performance in a specified area of performance. It seems unlikely to us that the Coast Guard with its preference for intuition will choose to go in that direction at the Academy. You might wonder why we are about to report research on the relationship between biographical data and performance if that is the way we feel. As students of organizational behavior our focus is on what we learn about the way an organization works by looking closely at who succeeds in that organization. The Coast Guard may or may not "tighten up" its screening of applicants based on their histories. We expect to trace the fortunes of the class of 1986 through their careers in order, in part, to see who succeeds. More importantly for us we hope to see what pattern of attributes is associated with success at different stages in an officer's career. We expect to learn something about the way we evaluate, train and assign in order that the Coast Guard may better educate and develop those who choose to serve in our organization. What follows is a report of a pilot study conducted on a subset of the class of 1986.

METHOD

The class of 1986

In July of 1982 250 cadets of the class of 1986 enrolled at the Coast Guard Academy. Four years later 142 graduated including 10 foreign nationals. Of the cadets who entered 18.3% were women. Twelve women graduated. The average math and verbal SAT scores of those entering was 643 and 559 respectively; 93% were in the top 20% of their high-school class and 88% were either 17 or 18 years of age. Following graduation all U. S. graduates were assigned to ships for initial tours of two years. For this study 48 graduates were selected at random. To these we added those graduates who failed to be selected for Lieutenant (Junior Grade) or LTJG for a total of 54.

Biographical data

The biographical data for each cadet was obtained from the standard candidate application. Each candidate filled out a standard set of documents although some applicants included extra letters of recommendation, newspaper clippings, etc. Most of the forms were those also in use at the other academies--the candidate personal data record (DD 1867) even had USAF Academy printed at the bottom. A list of the forms and a sample of their contents is given below:

1. Request for secondary school transcript-data re: high school size, percent attending college, students GPA and class rank with transcript attached.
2. Cadet candidate questionnaire-preferred major, four essays re reasons for interest in the Coast Guard and the Academy, race, sex.
3. School officials' evaluations of candidate-Three (coach, math teacher and English teacher) 10 item 6-point scale re ability to get along with others, gain respect etc. and short narrative.
4. Candidate personal data record-address, parents names and address(es), military history of parent(s), legal history, sports participation, military and educational history, extracurricular activities e.g., clubs, scouts, awards, CAP, ROTC.
5. Either ACT or SAT reports-sometimes both and often more than one report.

These data were coded for analysis along with some selected items that were derived as secondary data points from the forms. For example, the candidate's self reported grades listed on the SAT or ACT forms were compared with actual grades as reported on the

transcript. We did not code some data. For example, we did not code the school officials' scale scores due to a degree of range restriction which might serve to define the term. A distinction was made also in the record among those who submitted extra documents and those who did not.

Performance

Class rank at graduation was obtained for graduates. Two years following graduation the official performance reports of the graduated officers were reviewed. Even within the first two years it was obvious that neither an average of performance evaluations nor any single report gave an accurate picture of the officer's learning, assimilation and contribution. By the end of the second year each officer had established a clear and consistent record of performance; but, during those two years officers performed very differently. Some were identified as excellent performers from the start; others seemed either to take some time to "learn the ropes" or shed bad habits before they were evaluated as good performers. Few(six of 54) either were identified as "rising stars" or performed poorly enough to fail for selection to LTJG(also six out of 54). No one report gave a clear picture of an officer's performance;yet, each added to the picture. Therefore, we developed a composite score on the following scale:

5. Remarks indicate he/she is a special person--rising star, one of a kind in my career, etc.

4.. Excellent performer from the start but not identified as a star.

3. Either starts slow or has some reported weaknesses but has the confidence of the CO and is recommended for command.

2. After two years CO still isn't sure if this officer will make it, has some weaknesses that limit performance.

1. Failed of selection.

The mean performance score of the sample was 3.22 with the following distribution: 5, n=6; 4, n=18; 3, n=18; 2, n=6; 1, n=6.

RESULTS

There are in this small sample few variables that are related to performance following graduation. In order to compare high and low performers we combined those officers in the lowest and highest two rating categories leaving out the complex middle category for this

analysis (low= n of 12; high= n of 24). A t-test of high school performance and overall Academy performance showed that high school performance is not related to officer performance while Academy performance is (mean ranks -high performers 61.1 and low performers 88.3, $t=2.33$, $p<.05$). This seems reasonable but may surprise those who see the Academy as irrelevant to a Coast Guard career.

The only biographical data that were clearly related to officer performance were membership in the National Honor Society and leadership positions held e.g. team captain, class officer, or student government officer. Chi-square tests for these variables were 4.63, $p<.05$ and 3.85, $p<.05$ respectively(see Table 1) One other interesting note was the absence among all subjects of experience with the sea before reporting to the Academy. Only three had sailed before while two had been sea cadets.

TABLE 1
The Relationship between
Performance and National Honor Society
Membership and Leadership Positions

		low perf	high perf
Member	no	8	7
	yes	4	17
Leader ship	no	7	6
	yes	5	18

DISCUSSION

The results of this pilot study suggest that the Coast Guard should look more closely at indicators of prior leadership. Much of

the data that is now available to admissions personnel appears to be irrelevant with respect to future performance. For example, work history, either the fact of having a job or the number of hours worked, has no relationship with either graduation from the Academy or performance after graduation. It remains, however, important to select a "balanced" Academy class; and therefore, many factors which do not increase the cost effectiveness of educating Coast Guard officers will continue to influence the admissions process.

Even with this emphasis, however, there is still room and need to focus more attention on whom we select and graduate. A brief biography which we discovered as we followed individual records through the process of application, education and service may serve to highlight this need and opportunity. The CCEB reviewed his record and gave a strong recommendation for admission--he played more than one sport, stood in the top 10 of a large high-school class, had good verbal SAT scores and a 630 math score. Only one teacher said that he was immature and took the wrong things seriously. He was admitted. At the Academy he had the nickname "crazy dog;" in an interview before graduation he discussed his principles for deciding which regulations he kept--he broke a significant number. In general, whatever did not hurt someone else as he saw it was "OK." After graduation his Commanding Officer noted that he had "a temper" and sometimes verbally abused enlisted persons. His attention to his appearance was less than the CO desired. He was selected for LTJG and described as an excellent performer by the end of his first assignment.

Clearly this officer demonstrates a pattern of behavior that has been dimly evident in the record since high school. The Coast Guard has to decide whether these values, and behaviors are what it wants and, if not, determine whether the solution is an education/training issue, a performance evaluation and development issue or a selection issue. Perhaps in the long run we will decide that it is all three.

Predicting Leadership at the Service Academies and Beyond:
Discussant Remarks

Martin F. Wiskoff
Defense Personnel Security Research and Education Center
Monterey, CA

Background

A recent review of commissioning sources and selection criteria for military officers by Dianne Brown (1987) found that one of the major criteria for selecting among applicants to officer programs is leadership ability and potential. In her report, she discusses specific selection measures used by the various sources of officer commissioning. In relation to the procedures employed by the service academies, the term "whole-person" or "whole-candidate" is used to reflect that a combination of academic aptitude, physical ability and leadership potential predictors are employed to make the selection. For West Point, leadership potential is 30% of the "whole-candidate" score and is assessed through evaluation of an applicant's athletic participation and competence, and high school faculty recommendations. The other Academies similarly use combinations of academic indicators and other non-cognitive measures, presumably including estimates of leadership potential.

The relationships between academic predictors (H.S. performance and aptitude tests) and Academy performance criteria such as grades or class standing are understandable and easily interpreted. However, data relating leadership potential measures to subsequent leadership performance are not so clearly understood. Skeptics might conclude from reviews of the literature that we are searching for relationships that just don't exist. It is to the credit of the researchers on the panel that they chose to tackle such a thorny subject, but one which has extremely important operational implications for our military services.

The fundamental question, it seems to me, is whether it is possible to select personnel on "leadership potential" or whether the best we can do is to select those whom we think will be capable officers, i.e. the high quality, good guy approach, and then train them in leadership behavior.

Two of the papers on this panel addressed Naval Academy midshipman and leadership potential. This work was really born out of a 1984 decree from the incumbent Secretary of the Navy, John Lehman, that the Navy pay greater attention to graduating leaders from the Naval Academy. His notion was to decrease the emphasis on selection tests, move away from engineering and science as the preferred major and toward liberal arts and allow people to rise to their own levels, in a sort of open admissions system. The uniformed Navy, particularly the nuclear community, found Secretary Lehman's de-emphasis of scientific expertise distressing to say the least.

In 1984 I briefed the Assistant Secretary of the Navy for

Manpower, Chase Untermeyer, and proposed a compromise solution. My proposal, which was adopted, initiated research into selecting officers with leadership potential, while maintaining for the most part the existing Academy selection system. My hope was to introduce an evolutionary approach to influencing the characteristics of future generations of Navy officers rather than effecting a revolution such as Secretary Lehman was advocating. How successful has research been in developing these predictors of leadership potential at the Naval Academy?

The papers by Drs. Larry Stricker and Craig Russell reflect a concerted effort by the Academy, with the support of NPRDC, to investigate aspects of leadership during Midshipman status and subsequent to commissioning as an officer. NPRDC established the parameters of the program after in-depth review of (1) the current USNA selection system to identify potential predictors and criteria; (2) the leadership literature with a focus on the military literature; (3) other Academy procedures; and (4) results of a USNA meeting of military and leadership research experts.

Some of the factors that emerged from the planning stage were the: (a) potential of the Military Quality Performance Rating as an indicator of officer leadership potential and performance at the Academy; (b) development of a conceptual model focusing research efforts into six areas that have demonstrated a consistent relationship to leadership, and are also potentially assessable in 17 year-olds. These areas are intelligence, personality traits, task knowledge, interests, energy level, and social skills/styles); and (c) establishment of a research program supporting Naval Academy staff, Drs. Stricker and Russell, and work at the State University of New York, Binghamton with Dr. Bernard Bass on transformational leadership and the impact on subordinate motivation, team effects and superiors' appraisals.

The third paper, by CDR Earl Potter did not really address the issue of leadership as much as the development of personnel and success at the Coast Guard Academy, i.e. the good guy approach. As I did not receive the paper, I will not be able to comment.

Comments on paper by Dr. Lawrence Stricker, "Assessing Leadership Potential at the Naval Academy with a Biographical Measure."

Dr. Stricker draws support for his research from a recent assessment by Hogan (1987) which suggested the predictiveness of leadership from personality tests. He also noted that this contradicts earlier respected reviews by Stogdill (1948) and by Gibb (1954) which found just the opposite, i.e. leadership was situational.

Dr. Stricker uses a construct approach to developing predictors. He selected five personality traits from a literature review that offered some support for being able to predict leadership. The five scales, plus a Social Desirability scale, were carefully developed, pilot tested and reviewed by

several panels. Reasonable alphas were obtained for the most part.

It should be noted that predictor data were obtained on Midshipman at the Academy during Plebe summer so that unrestricted data on applicants were not available. The criteria used reflect both the more specific leadership measure and general officer "goodness". For the former, peer ratings obtained on the last weekend of Plebe Summer are a very appropriate measure of leadership. I share Larry Stricker's doubt about the secondary criteria of Military Performance and Professional/Military Quality Point Rating (MQPR) as leadership indices. However MQPR does include peer ratings as one of its components.

A few comments on the findings. It is unclear why the social desirability scale has such low reliability (.27) and why some of the highest intercorrelations of scales were with Social Desirability despite the low reliability.

The paper, probably because of length restrictions did not contain intercorrelations among the criteria. However, material sent to me by Dr. Russell included an intercorrelation matrix which showed that peer ratings of leadership correlate .41 with MQPR and .45 with one component of MQPR, the grade assigned by an individual's company officer based on performance assessments by peers and senior officers. Peer ratings correlated only .07 with Academic Quality Point Ratio, the individual's academic grade point average at the end of his/her first semester at the Academy.

Recognizing that this is a preliminary study I perhaps shouldn't criticize the lack of cross-validity. However given the sizable N's it would have been possible to evaluate how well the predictors hold up by crossing back to the item analysis sample of 233 1987 high school graduates or using the sample of 171 earlier graduates.

I was encouraged by the relationship between the Sociability scale and Leadership ratings (.28). However, if there was a halo effect operating, as suggested by Dr. Stricker, it is no surprise that the personality trait of sociability is most highly related to the halo effect.

Overall I found it a very well-done study, with promising findings, and certainly worthy of follow-on effort. I look forward to the results that will be obtained on applicants to the Academy and the determinations of incremental validity to currently used measures

Comments on paper by Dr. Craig Russell, "Predicting Leadership at the Service Academies."

Craig Russell's approach to using biodata is embedded in quite a different theoretical framework. Dr. Stricker is looking for enduring traits, Dr. Russell for life experiences, i.e. lessons learned by the individual that shapes future behavior.

Dr. Russell makes the point that we need to define military leadership and its developmental stages. In his

written paper he states "the transformational Naval leader would communicate and inspire goals reflective of the values, beliefs and needs of the country and manage subordinate activities relative to that goal." The use of the concepts leadership and management in the same context is one that Craig Russell explicated quite clearly.

Let me digress for a moment from Dr. Russell's paper to bring in some information from a conference that was held at the Naval Academy in June 1987 entitled, "Military Leadership: Tradition and Future Trends." I will draw from Dr. Wally Sinaiko's (1987) excellent minutes of the meeting.

Over 90 uniformed officers and civilians attended and there were 35 speakers. There were a dozen flag officers (active and retired), including the Chief of Naval Operations, and a past chairman of the Joint Chiefs of Staff and at least a half-dozen academic researchers noted for their contributions to the field of leadership.

In terms of defining leadership, as you can imagine there was quite a diversity of opinion. To the question, "Are leaders born or can they be made?", respondents came down on both sides. Interestingly, Bernard Dodd, a Royal Navy Senior Psychologists said that the military in the United Kingdom does not address the concept of leadership, rather it is inferred.

The issue of management vs. leadership was addressed by many participants. It was generally felt that they are "complimentary" but do not necessarily exist in the same individual, i.e. a leader focuses on people, a manager on activities. I think we must be careful concerning our terminology if we are ever to have any hope of operationally defining leadership.

Returning to Dr. Russell's paper, he was kind enough to send me copies of his writings entitled the "Theory and practice in the selection and development of organizational leaders" and the "Predictive validity of biodata items generated from retrospective life experience essays." I find his research to be extremely programmatic and well-thought through. I appreciate his attempts to define what he means by a "fully developed" leader as a ship's captain. It would be nice to establish the dimensions of junior officer leadership and then develop them into measurable criteria.

The training implications of Dr. Russell's work are intriguing. As reported by CAPT Douglas Katz at the Naval Academy Leadership Conference I mentioned earlier, one-third of the total number of credit hours at Annapolis have to do with professional development education. Similarly, LCOL Robert Gregory said that throughout the four years at the Air Force Academy there is a lot of emphasis on feedback to students on their leadership performance. COL Howard Prince of the U. S. Military Academy indicated that both formal study and observational learning are used in training leadership. West Point provides remedial training for those who are deficient in leadership skills.

Dr. Russell proposes that we identify developmental stages for midshipmen or cadets and tailor the training accordingly.

I'm not sure of the feasibility or practicality of this notion, but it certainly is food-for-thought and empirical evaluation.

I found Dr. Russell's paper and written material to be stimulating and full of testable hypotheses. As with Dr. Stricker, I will await the results of their analyses on Academy applicants.

Concluding Remarks

Where do we stand in our knowledge of military leadership, particularly as a precursor to its prediction? Some impressions may be obtained from the 1987 Naval Academy conference since it was such a significant assembly of senior researchers and policy makers. Let me paraphrase some of the summary remarks made by Professor David Segal of the University of Maryland; a military sociologist who currently is on the staff of West Point for a one-year period:

1. The field of leadership has moved away from trait theory as its primary explanatory mechanism.

2. Leadership is not innate, however the service academies place too much emphasis on leadership development, with a consensus about how to train leaders, but little agreement on what "leadership" means.

3. Leadership and management are distinct behaviors, although there is no consensus about what each one means.

4. There are cross-national differences in leadership styles and also ethnic differences within nations.

5. Leadership is a family of processes and it has no single correct style.

6. The concept of "charisma" is more prevalent these days as perhaps seen in the emergence of "transformational" leadership styles and explanations.

7. The academies train only a small percentage of military leaders but little attention is being given to the other officer sources such as ROTC and OCS.

I have three additional remarks I would like to make related to research in Academy settings, the use of biodata and the study of leadership.

A major feature of the Academy research programs has been their completeness in looking at the "whole-person". To a great extent, as researchers we have been blessed with a fertile research environment, willing consumers and a rich source and variety of data. I am pleased to see that the programs are expanding in two directions: (a) more conceptual bases for the work rather than dust bowl empiricism and (b) extending the behavioral horizon in both directions, i.e. tapping earlier life experiences and measuring later career behavior. I wish we would have had the benefit of descriptions of research at West Point and the Air Force Academy to round out our picture.

We have to realistically address the predictive limitations of biodata, whether we base our work on traits or other underpinnings such as life experiences. Just as we have come to expect a maximum correlation of .55-.60 for academic

prediction, an r of .30-.35 seems to be a reasonable target for biodata in predicting criteria such as peer ratings and other performance assessments. My point is that if you employ a variety of non-cognitive predictors and regress them against rating indices of success, there usually will be several combinations of predictors that achieve the same multiple R . Given this, the more crucial decision is the practical implication of using the given predictors on the applicant population to meet the policy goals of the institution. A case in point was the example I gave of Secretary Lehman's policy changes to Academy selection programs. It is interesting to note that the Academy is going back to focusing on engineering and science for selecting applicants now that he no longer is directing Navy policy.

My third comment is that I am not convinced that we understand leadership dynamics sufficiently to be able to predict potential in applicants to the Academies. I believe we can obtain r 's up to .30 with surrogate measures of leadership rating, but we have a long way to go to interpret the findings. Our goals should be quite modest in adding incremental validity to the powerful selection systems that operate at the Academies. We have played a large role in the development of these systems and their data bases. We should strive to clarify the concept of military leadership to enable improved systems for selecting and training officers to be developed.

References

Brown, D. C. (1987). Military officers: Commissioning sources and selection criteria. HumRRO Final Report 87-42. Alexandria, VA.

Sinaiko, H. W. (1987) Military leadership: Traditions and future trends (unpublished minutes). Smithsonian Institution, Alexandria, VA

Disclaimer

The opinions expressed in this paper are those of the author, are not official, and do not necessarily reflect the views of the Department of Defense or the Federal Government.

Medical Standards for Enlistment and The Qualified Military Available Population

Michael T. Laurence
Defense Manpower Data Center

Historically, most of the research conducted by the Department of Defense in the area of military enlistment standards has concentrated on the aptitude and education criteria. Indeed, when most manpower analysts speak of recruit "quality" they are usually referring to those recruits who are high school graduates and fall in AFQT Categories I-IIIa. In these terms, little attention is paid to the roles of the moral character or medical fitness criteria in defining quality.

In view of the decline in the size of the military-age youth population and the continuing needs of the Military Services for quality manpower, obtaining information about medical fitness has become of more immediate interest. In 1984, the Defense Manpower Data Center (DMDC), under the auspices of the Directorate for Accession Policy, Office of the Assistant Secretary of Defense (Force Management and Personnel) responded to this interest and initiated a research program in the area of medical fitness standards. The goals of this program were to provide data describing the size of the available manpower pool under the medical fitness criterion; develop and assess methodology for establishing medical standards for enlistment; and provide data relating medical fitness to military performance measures.

This research program began in the traditional way with a review of the literature which yielded two significant findings. First, surprisingly little work has been completed in the area of medical standards. The second result was that little of the published work is of utility to present day manpower planners and analysts. For the most part the literature focused on the presentation of historical descriptive statistics that summarized the medical fitness of those individuals who either applied for military service or were, prior to 1974, conscripted. Collected from applicants and draftees, these data are not appropriate for describing the medical fitness of the contemporary American military-age cohort. The data are not representative of all American youth due to self-selection among applicants, the vagaries of the draft, and the unquantifiable effects of the changing social, economic, and political climates over time. The data are further biased by the absence of medical fitness data for those applicants who were determined to be unsuitable as a result of ad hoc initial screening by recruiters and formal screening by the use of aptitude test results.

The only noteworthy study was followed, and one which contains the defects just described but at the same time serves as a model for our present work, is the 1964 report of the President's Task Force on Manpower Conservation entitled One-Third of a Nation. This report was the only major effort that used actual examination data to specify medical fitness standards for all

Paper presented at the 1985 Annual Conference of the Military Testing Association, Washington, DC, November 1985. The views expressed in this paper are those of the author and do not reflect the policy or position of the Department of Defense.

enlistment criteria then in effect. The principal finding of this study was that one-third of all men turning 18 would be found unqualified if they were examined for induction into the military. More precisely, by generalizing the examination results of applicants and draftees between August 1958 and June 1960 to the general population, it was estimated that 31.7 percent were not qualified for military service. Over half of all those found not qualified (16.3 percent of all 18 year-olds) were disqualified for medical reasons.

In the absence of more recent, unbiased and representative data when asked about the general disqualification rate among applicants for military service, we relied on these statistics. Clearly, this is less than ideal. It is unlikely that the disqualification rate has not changed in the thirty years or so since the data upon which it is based were collected. It is even more unlikely that they accurately describe the medical fitness for military service of the general population. Finally, only statistics for males were reported.

In the course of our work we identified an alternative to the use of the historical data bases in the form of the National Health and Nutrition Examination Surveys (NHANES). Conducted by the National Center for Health Statistics, these surveys provide a comprehensive and representative set of medical history and physical examination data for national probability samples of military-age American males and females that were ideal for our analytic needs. The utility of the NHANES data bases has been demonstrated by DMDC's determination of the overall medical disqualification rates and our studies of enlistment height and weight standards. For the first time, we have determined the percentage of all military-age men and women who are medically qualified for military service, how the modification of a medical standard would affect the supply of manpower, and the relationship between medical fitness and military performance as measured by attrition.

In 1986, we published a study entitled The Medical Fitness of American Youth for Military Service (Overbey, Winter, & Laurence, 1986). In this study we took the NHANES II medical examination data of the 1,339 males and 1,370 females between the ages of 16-24 years and assessed their fitness for military service as prescribed in DoD Directive 6130.3 and under the Army height and weight enlistment standards. As shown on the first line of Table 1 we found that 18.2 percent of males and 41.1 percent of the females were medically disqualified from serving in the military. The leading cause for disqualification among both males (16.9 percent of the total) and females (67.4 percent of the total) was overweightedness.

The 18.3 percent disqualification rate for males appears to be only marginally higher than the 16.3 percent rate reported in One-Third of a Nation. However, the similarity of these overall disqualification rates masks real underlying differences. A comparison of the previously reported prevalences of disqualifying causes to those found in the our study revealed wide variations which are likely due to the non-representativeness of the previously reported data.

Since the NHANES II did not include data for all the medical criteria, the results in Table 1 understate the true disqualification rates among the population of military-age Americans. Adjustments to the disqualification rates for visual acuity and limitation of motion is likely to increase the

Table 1

**Medical Disqualification Rates Among 16-24 Year-Old Americans
For All Causes and For Overweightness**

	<u>All Causes</u>		<u>Overweight</u>	
	<u>Males</u>	<u>Females</u>	<u>Males</u>	<u>Females</u>
Medical Fitness of American Youth	18.3	41.4	3.1	27.9
Adjustment for Missing Criteria	3.4	3.0	-	-
	21.7	44.4	3.1	27.9
Adjustment to 135% BM Standards	(.7)	(22.5)	(.2)	(22.7)
	21.0	21.9	2.9	5.2
Adjustment to 120% BM Standards	4.0	5.7	4.0	5.7
Revised Estimate	25.0	27.6	6.9	10.9

overall disqualification rates among the population of 16-24 year-old males and females to approximately 21.7 percent and 44.4 percent, respectively.

As noted earlier, the large difference between the overall male and female disqualification rates is due to the huge difference in the prevalence of overweightness. A small part of the difference can be accounted for by the fact that, within the general population, the prevalence of overweightness among females is higher than among males. Not only is this fact reported in the medical literature, it was supported by the NHANES II data. Most of the difference in the disqualification rates, however, was accounted for by the structure of the height and weight standards themselves (see Laurence, 1985). A comparison of current maximum weight standards to the mean body-weights of the population revealed that the maximum weight standards for females, as a percentage of the mean body-weights, were consistently lower than those for males. In other words, the maximum weight standards for females were much more restrictive than those for males.

The next step in our analysis was the development of weight standards for females that were methodologically consistent with the male standards. This development of alternative standards employed the use of an analytic tool, the body-mass index, that permits the simultaneous evaluation of height and body weight. The analysis of the current standards for young males revealed that the maximum weights were set at 135 percent of mean body-mass. Accordingly, we determined the mean body-mass for females and then applied the formula used for the males to produce comparable maximum weight standards. The result is shown in Table 1 as the "Adjustment to 135% BM standards". The 21.8 percentage point difference between the male and female disqualification rates (3.1 percent versus 24.9 percent, respectively) under the current standards was reduced to 2.2 percentage points difference (2.9 percent for the males versus 5.2 percent for the females). This remaining difference in the qualification rates is due, in part, to the continuing prevalence of overweightness among males and females, a condition which, in turn, disquali-

fication rates were recalculated the disqualification rates for males (21.0 percent) and females (21.9 percent) were nearly the same.

Finally, in recognition of the fact that the maximum weight standards under the 135 percent of mean body-mass formula permitted some individuals who could be considered medically overweight to enlist, we created still another set of alternative weight standards using the medical definition of overweightedness as body weight 20 percent or more in excess of the mean. When these standards were applied the disqualification rate for males increased by 4.0 percentage points and the rate for females increased by 5.7 percentage points yielding overall disqualification rates of 25.0 percent for males and 27.6 percent for females.

Returning to the introductory theme of this paper, these data now permit a more complete description of the quality of available manpower in terms of aptitude, education, and medical fitness. Presented in Table 2 are combined estimates of the percentage of military-age Americans who are disqualified for each of the Services on these three criteria. For comparison, the results from One-Third of a Nation are also presented. The disqualification rates on the aptitude/education criteria are taken from Screening for Service (Eitelberg, Laurence, & Waters with Perelman, 1984). The medical disqualification rates are based on the results in Table 1 which have been further adjusted to reflect the height and weight standards actually used by each of the Services. (Subsequent to the publication of The Medical Fitness of American Youth the Navy implemented new standards that were more restrictive for males and both the Navy and Air Force implemented less restrictive standards for females.)

The total disqualification rates for males and females, regardless of Service, are all substantially higher than those reported for males in One-Third of a Nation. If One-Third of a Nation were published today and the data for males and females averaged together, the title would have to be revised to "Three-Fifths of a Nation." The higher disqualification rates of females, compared to males, are due to the higher aptitude/education standards applied to females by all the Services and, for the Army and Marine Corps, the more restrictive maximum weight standards applied to females. Aside from these differences, the most notable feature of Table 2 is the number of question marks that indicate unknown percentages. Were disqualification rates for the moral character criteria known and added to the totals, the combined disqualification rates would be even higher. On the other hand, were accurate esti-

Table 2
Estimates of the Percentage of Military-age American Youth
Disqualified for Service on the Various Enlistment Criteria

	Males					Females				
	<u>One-Third of a Nation</u>	<u>Army</u>	<u>Navy</u>	<u>Marine Corps</u>	<u>Air Force</u>	<u>One-Third of a Nation</u>	<u>Army</u>	<u>Navy</u>	<u>Marine Corps</u>	<u>Air Force</u>
Aptitude/Education	13.0	23.7	25.0	27.6	37.4	?	21.7	42.2	53.6	59.6
Medical	16.3	21.7	24.1	21.7	28.3	?	44.4	24.4	39.8	33.0
Moral	3.9	?	?	?	?	?	?	?	?	?
Both Aptitude and Medical	(1.5)	(?)	(?)	(?)	(?)	(?)	(?)	(?)	(?)	(?)
All Other Combinations	(?)	(?)	(?)	(?)	(?)	(?)	(?)	(?)	(?)	(?)
Total Disqualified	<u>31.2</u>	<u>45.4</u>	<u>49.1</u>	<u>49.3</u>	<u>65.7</u>	<u>?</u>	<u>66.1</u>	<u>66.6</u>	<u>93.4</u>	<u>92.6</u>

mates of the percentages disqualified on both the aptitude/education and medical fitness criteria available, as in One-Third of a Nation, as well as all other combinations of multiple disqualification, the combined disqualification would be reduced. Clearly, there is room for improvement in our attempt to specify the qualified military available population. Towards this end, we are assessing the feasibility of a comprehensive study in which a nationally representative sample of military-age Americans would be subjected to examination on all the enlistment criteria.

Next, I would like to describe our work relating medical fitness to military performance. In the discussion of the percentage medically qualified data in Table 1 the effect of weight standards based on the medical definition of overweightedness was illustrated. Implicit in this discussion was the notion that military performance and body-weight might be related. Recently we completed a study (Laurence, 1988) that directly addressed this notion.

Using basic training and 36-month attrition as the criterion for successful performance, this study compared the attrition rates of FY 1983 accessions who could be characterized as overweight under the then current accession standards and the medical definition (the 120 percent of mean body-mass standards) and those who were not overweight. As shown in Table 3, the attrition rates for overweight males and females under the current standards were consistently higher than the rates for those not overweight. Not surprisingly, when the more restrictive 120 percent of mean body-mass standards were applied the attrition rates for overweight males were even higher. In contrast, application of the 120 percent of mean body-mass standards to the females resulted in substantial decreases in the attrition rate among those found overweight. This result is due to the fact that under the restrictive current accession standards very few females were accessed who could be characterized as overweight under the less restrictive 120 percent of mean body-mass standards. Given this situation, there were very few medically overweight females who might attrit.

These results clearly indicate that the attrition rates of males could be reduced by employing more restrictive accession weight standards. However, had the 120 percent of mean body-mass standards been in effect, and strictly adhered to, when the FY 1983 male accession cohort was recruited, 15,396 of the actual accessions would have been excluded and would have had to be replaced with other qualified recruits. But, given the lower attrition rates for those qualified under the 120 percent of mean body-mass standards, the

Table 3
DoD Basic Training and 36-Month Attrition Rates
Among Not Overweight and Overweight FY 1983 Non-Prior Service Accessions

	Basic Training Attrition				36-Month Attrition			
	Males		Females		Males		Females	
	Current Accession	120% Mean Body Mass	Current Accession	120% Mean Body Mass	Current Accession	120% Mean Body Mass	Current Accession	120% Mean Body Mass
Not Overweight	8.4	8.1	10.6	10.8	26.9	20.6	33.4	34.0
Overweight	12.1	14.2	12.9	3.4	33.8	34.8	30.4	25.4
Percentage Change	+46.0	+75.1	+21.7	-68.5	+24.0	+15.9	-18.0	-25.3

15,396 excluded accessions could be replaced by 13,835 qualified recruits to yield the same number of non-attritees at the end of the 36-month follow-up period.

In broadly stated terms, we have embarked on a research program that will match the breadth and scope of the program evaluating and assessing aptitude standards for enlistment. Medical fitness standards have for the most part remained unchanged since the end of World War II partially by design (e.g. the maximum standards have been prescribed by law) and partially by default (e.g. there has been no compelling reason to change them due to manpower shortages). In addition, there has been a real lack of empirical data and information upon which to propose and evaluate changes.

The current standards of medical fitness are predicated on the requirement that each and every accession meet the same high standard of fitness for combat anywhere in the world without regard to their eventual assignment. This requirement comes out of an era in which many accessions (e.g., draftees) had little choice as to their eventual military occupation. This is a "luxury" that in the past may have been easily afforded but one which, in view of the decline in the size of the manpower pool may become difficult to sustain. Just as applicants are currently evaluated on the aptitude criteria and placed in categories which are matched to a particular military occupation, a similar concept applied to medical fitness standards may be one whose time will come. In any event, effective policy formulation requires an adequate understanding and grasp of the complexities of the issue in question. To make good policy in the area of medical fitness standards requires an understanding of the effect of current policy and the likely effects any changes might produce.

References

- Eitelberg, M.J., Laurence, J.H., & Waters, B.K. with Perelman, L.S. (September 1984). Screening for Service: Aptitude and Education Criteria for Military Entry. Washington, DC: Office of the Assistant Secretary of Defense (Manpower, Installations, and Logistics).
- Laurence, M.T. (April 1988). Enlistment Height/Weight Standards and Attrition from the Military. Arlington, VA: Defense Manpower Data Center.
- Laurence, M.T. (November 1985). Development of a Methodology for Establishing Joint Service Height and Weight Standards for Enlistment. Arlington, VA: Defense Manpower Data Center.
- Overbey, J.W., Winter, P.E., & Laurence, M.T. (September 1987). The Medical Fitness of American Youth for Military Service. Arlington, VA: Defense Manpower Data Center.
- The President's Task Force on Manpower Conservation. (January 1964). One-Third of a Nation: A Report on Young Men Found Unqualified for Military Service. Washington, DC: U.S. Government Printing Office.

CHILDREN OF MILITARY: FORGOTTEN COMBAT MULTIPLIERS?

Michael E. Freville, Ed.D., and Rose Webb Brooks, M.A.
Fort Knox Community Schools, Fort Knox, Kentucky 40121

The title of our paper may be confusing at least and perhaps shocking at most, yet our purpose is to provoke your thoughts today, and share with you the culmination of 22 years of experience Mrs. Brooks and I have dealing with so called "Army Brats" (and we use that term affectionately) at Fort Knox, Kentucky. We believe our observations, conclusions, and recommendations can be generalized across the various services and apply to all military families.

I want to talk with you about something that is ultimately very important to us all - combat readiness. You may wonder what that term has to do with children. It has been our observation that military service member's overall ability to train efficiently and purposefully in peacetime, to be deployable for extended TDY (temporary duty) training at other locations, whether in CONUS (Continental United States) or overseas, and, ultimately, to be combat ready when the so-called balloon goes up, is very, very much dependent on how that service member perceives his/her children to be "OK". What do we mean by "OK"? Let's look at some of the key factors that we believe influence the service member's readiness posture.

As educators, we have seen first-hand the importance of easily accessible in-school counseling services for kids. You know, before I came to Fort Knox I had the pre-conceived notion or stereotype that kids of military parents were perfectly behaved little soldiers. WRONG! They're just like all other kids, except for two important differences: they move frequently, and their moms or dads are in the business of training daily for war, some more so than others, such as the ones in the combat arms. These two differences add a significant amount of stress on young people who are already going through the normal stresses and pressures of growing up which, believe me are much more than when I grew up. When dad or mom comes home after a long duty day, they don't need the further stress of knowing that their kids are not doing well in school, having trouble adapting to a new school and community, and having peer group problems, along with other pressures. We deal with many service members who can't possibly be giving anywhere near 100% on the job because they may perceive that their kids are not receiving proper educational and psychological services. Whenever anyone is under stress of any type, including the question of whether or not his or her children's needs are being met, then it most certainly follows that school can and will be affected.

We have seen the positive and potential influence and effects of having on-post schools staffed with people who are sensitive to the special needs of military kids. We have students move in and out almost daily, which is certainly stressful on those who

move, and those who watch others move, losing friends, and knowing they too will be moving - next month or next year. We deal with parents on a daily basis who are having trouble with their kids and need immediate help. You know, people who live on military reservations are accustomed to having facilities close by - the PX, commissary, hospital, etc. It is our belief that there should be schools there also, and the reason is that military kids who go to the same school draw support and strength from each other. Additionally, the service member knows that the school is easily accessible and the adults in the building are sensitive to the uniqueness of the military child. This is just plain not true of most public schools who are not accustomed to the unique lifestyle of the military population.

Obviously, as counselors, we believe we play key roles in the emotional stability of students, and, believe me, we do it every day. Since we play that key role, and because kids who are experiencing the normal and unique problems and pressures of growing up in a military environment and who receive help in a sensitive manner, it follows suit that the service member's concept that his/her kids are "OK" will remain intact. I don't mean to make this simplistic, but this has been our direct experience and what military members have told us. We are proponents of schools located on military reservations because this proximity factor seems critical. Physical and mental proximity is critical because the parent must know that the school is physically close. After all, all the other services his/her family receives are physically close. It must be mentally close, and by this I mean a comfortable feeling of knowing the school truly accepts and understands the child.

Additionally, we are proponents of military hospitals having uniformed or civilian psychologists available for family members. It has been our experience at Fort Knox that this isn't true. CHAMPUS (Civilian Health and Medical Program for the Uniformed Services) is fine for some families, but many tell me that either they can't afford their share of the cost, or they won't use it because it means going off-post and; therefore the proximity factor is lost. We, as school counselors, do all we can to help the emotional stability of the family, but our caseloads are extremely high, and we are not equipped to handle on-going therapy.

In summary, military and combat readiness is a multi-faceted concept. It is a complex concept. It is composed of weapons systems, training, funding, more training, allocation of time and human resources, and other "hard" factors. But we submit for your consideration that a crucial component of that readiness is mental attitude - the assurance the service member has that his/her child and family are stable, and receiving proper care both educationally and emotionally. We submit to you the need for on-post schools and increased psychological services for family members in military hospitals. We owe the soldier, the seaman, and the airman the assurance that when

he/she deploys, they're not going off to one battle and leaving another raging at home. In war, there is nothing second in importance to readiness!

THE IMPACT OF SPOUSE EMPLOYMENT ON MILITARY MANPOWER¹

Paul A. Gade, Newell Kent Eaton, and Roya Bauman, U.S. Army Research Institute for the Behavioral and Social Sciences, Alexandria, VA

Over the next decade more military spouses than ever before will want to work outside the home. The number of enlisted wives entering the labor force has nearly doubled in the last nine years. Greater desire for employment in military spouses has lead to greater unemployment for many military spouses even though spouse employment programs have expanded. Unemployment for junior enlisted spouses, for example, was 45% in 1987. We project that spouse unemployment at the these currently high rates will have a very deleterious effect on retention as the need for two incomes increases in American society. The military services will need to develop cost-effective strategies for combating the spouse unemployment problem. One solution to the spouse employment problem, we suggest, might be to stabilize families in one location for a longer period of time so that spouses will have a better chance of finding employment commensurate with their skills. Separations such as TDY and short unaccompanied tours may then become more acceptable to families provided that PCS moves are less frequent.

CHANGING DEMOGRAPHICS

The national trend is toward dual income families in the United States. More than half of the wives in American families now are employed outside the home and some projections show that by 2000, 80% of all wives will be in the work force (Harris, 1987). This means that the dual income household has become the "norm" in American society. Are the military services following this trend we see in the general population? Results from the Current Population Survey show a dramatic rise in the number of military wives in the labor force during the last couple of years. In 1985, 67% of non-military wives in the 16 - 45 year old age-group were in the labor force as were 52% of military wives. In 1987, 70% of the civilian wives were in the labor force, while the percentage of military wives jumped to 60%. This increase in labor force participation is not equal for officers and enlisted wives, at least within the Army. The percentage of Army officer wives in the labor force has increased only slightly over the last eight years, moving from 52% in 1979, to 54% in 1985 to 57% in 1987. The percentage of enlisted Army wives who are working or looking for work has risen dramatically from 44% in 1979 to 64% in 1987; rapidly approaching the labor force levels of non-military wives in the same age group!

¹The views expressed in this paper are solely those of the authors and do not reflect the views of the Department of the Army or the U.S. Army Research Institute for the Behavioral and Social Sciences.

Furthermore, more than half of these enlisted wives have children under the age of five, compared to 41% for non-military wives in the same age group.

Unfortunately, unemployment among Army enlisted wives has risen even more dramatically than has labor force participation, rising from a low of 7% in 1979 to 33% in 1987. The news is even worse for the youngest enlisted wives whose unemployment rate is a whopping 45% (Schwartz, 1988; Schwartz, Wood & Griffith, 1988; and Griffith, Doering, & Mahoney, 1986)!

Why are more of these wives working or looking for work? When Army wives were asked that question on the 1985 DoD survey of military spouses, the majority of the enlisted wives (59%) said they worked because they need the money. Officer wives, on the other hand, said they worked primarily because it gave them independence (54%) or simply because they enjoyed it (44%). Only 31% of the officer wives felt they need to work for the added income. Results from the DoD survey clearly show that enlisted families are dependent on spouse earnings in that 30% of family income for enlisted soldier families comes from spouse employment (Griffith, et al., 1986).

SPOUSE EMPLOYMENT IMPACTS THE CAREER DECISIONS OF SOLDIERS

Research has clearly established that spouse satisfaction with the Army has a direct impact on soldier reenlistment decision making. We believe that spouse employment is now one of the key factors in spouse satisfaction with the Army, and that it will take on increasing importance as dual earner families become the norm in the military services as they have become in the civilian population. We estimate that 20% to 30% of the soldiers who leave active duty make this career change largely for family-related reasons. Our research shows that next to the soldier him or herself, the most important influencer is the soldier's spouse, followed far behind by other family members. In a recent investigation, we found that for those soldiers who told us they were going to reenlist, 73% reported that their spouses had positive attitudes about the Army, while only 5% reported that their spouses were negative. This supports earlier data that showed that 83% of soldiers whose spouses were satisfied with military life intended to reenlist compared to only 49% of soldiers who reported that their spouses weren't very satisfied. Spouse support and satisfaction with the Army is clearly an important factor in the reenlistment decision the soldier makes. Analyses of the 1985 DoD survey have established that spouse satisfaction with the Army can be directly affected by the ability of the spouse to find a job they can be happy with. Our analyses show that the wives who reported that they were happiest with the Army were those who were working in a job that was compatible with their husband's job. Wives who were unemployed were significantly lower in their satisfaction with the Army. However, the unhappiest wives were those who were working in a job that presented serious conflicts with their

husband's job. So it's not just a matter of finding wives jobs, we must also make sure those jobs don't present more work - family conflicts than they solve.

Even more direct evidence of the important influence of spouse employment on reenlistment decision making is found in some of our recent work with the Annualized Cost of Leaving Model or ACOL, as it is usually known. ACOL is a model that attempts to describe reenlistment behavior as a function of potential earnings inside the Army vs those outside the Army. Over simplifying a bit, if the soldier thinks he or she can earn more outside the Army over the next few years, he or she is not likely to reenlist. Army policy makers can use this model to predict what the effects of a pay raise or lack thereof would be on reenlistment rates in the Army. We have taken the model one step further by including a similar choice for spouse earnings and have found that a dollar increase in spouses earnings has the same effect on reenlistment intent as does a dollar increase in soldier earnings! This means that reenlistments can probably be increased by increasing spouse employment. Unfortunately, it also probably means that reenlistments will be lost if spouse unemployment continues at the high levels we see now. We expect this will be especially true for the best and the brightest of our soldiers, since it has been clearly shown that the desire for spouse employment is directly related to education level, and we have recruited bright, highly educated soldiers in large numbers during the past few years.

WHAT NEEDS TO BE DONE TO IMPROVE SPOUSE EMPLOYMENT?

Our analyses of the DoD survey results suggest that current spouse employment programs help to increase labor force participation and the subsequent employment of military spouses. When we compared installations with and without spouse employment programs, we found that installations with spouse employment programs had higher labor force participation (59% vs. 46%) while maintaining equal employment rates at 76%. Thus installations that had spouse employment programs had about 30% more spouses employed than did those installations without such programs. One must be cautious in interpreting these results, however, since it may be that installations with high employment possibilities encourage the implementation of spouse employment programs; while those with few local employment possibilities may see spouse employment programs as a futile effort.

Some of our analyses of the DoD survey show that Army spouse employment could be improved by stabilizing assignments. Our results clearly show that spouse employment likelihood improves monotonically with increased time at a particular location (Schwartz, et al., 1988). The quality of that employment also improves, but not so rapidly, with increased time at a particular location. Also, results for both quantity and quality of spouse employment as a function of tour length are not nearly so dramatic for those families stationed in Europe. Nevertheless,

the trends are clear. If you can stabilize families in one place for a longer period of time, you can increase spouse employment and perhaps thereby increase retention as well.

Spouse employment carries some special problems for the military as well. Since military working wives are also likely to be working mothers as well, child care becomes a major concern. In America's civilian families where both husband and wife work, child care is handled as follows: 25% leave their children with a relative -- usually a grandparent; 21% arrange their work schedules so that one of the parents is always at home to care for the children; 11% have a housekeeper at home; 10% take their children to a day-care center; 8% get a friend or neighbor to watch their children; 7% have an older sibling look after younger ones; 6% let the child care for themselves until one of the parents gets home from work; and 5% send their children off to a nursery school (Harris, 1987). We don't know yet what similar numbers are for Army families, but it's safe to say that the most frequently used child care method for civilians, leaving children with a relative, is probably not available to most Army couples. It's also more than likely the case that arranging work schedules so that one of the parents can be with the child is far more difficult in Army marriages -- especially in those marriages where both husband and wife are in the Army.

WHERE DO WE GO FROM HERE?

Since we believe that increased spouse employment in the military is inevitable, we need to know what the impacts of increased spouse employment are likely to be and how the military can adapt to them. We need to clearly establish the links between spouse employment and retention and readiness. For example, we think that spouse employment affects spouse satisfaction with Army life which in turn influences the soldier's intent to reenlist; however, there may be even more direct links than this such as the link between spouse earnings and retention intention we found in our revision of the ACOL Model. Spouse employment can also have an impact on readiness in several ways. For example, the need for soldiers to moonlight may be reduced by the financial relief that spouse employment provides. Soldiers' availability for duty will undoubtedly be adversely effected by spouse employment in much the same way it has in the civilian world. We need to know how this will effect readiness.

We suggest that stabilizing PCS moves may be a cost-effective way to increase spouse employment, at least for the Army. One way this might be accomplished is to home station families and move soldiers to overseas or remote site tours on periodic, short (e.g., 12 month), unaccompanied tours of duty. One would have to be careful in executing such a plan since soldiers perceive that the most deleterious thing for their marriages are separations from their spouses. Most of our

officers and enlisted dual Army career couples, for example, told us they wouldn't put up with a separation of a year or more. There is a bright note, however, in that wives report that one of the effective ways they cope with separations is to get a job. By working, they overcome boredom and busy themselves with meaningful tasks that keep them from dwelling on their separation. Spouse employment then is a very effective way to help soldiers and, in particular, their spouses to cope with separations caused by such short tours.

REFERENCES

- Griffith, J. D., Doering, Z. D., and Mahoney, B. S. (1986). Description of Spouses of Officers and Enlisted Personnel in the U.S. Armed Forces: 1985. Defense Manpower Data Center, Arlington, VA, November, 1986.
- Schwartz, J. B. (1987). Labor Force Participation, Employment, and Earnings of Married Women: A Comparison of Military and Civilian Wives. Draft report, Research Triangle Institute, October.
- Schwartz, B, Wood L. and Griffith, J. (1987). The Impact of Military Life on Spouse Labor Force Outcomes. Draft report, Research Triangle Institute.
- Harris, L. (1987). Inside America. New York: Vantage Books.

Contributions of Spouse Related Factors Affecting Reenlistment for Enlisted Personnel¹

Alfred L. Smith, Jr.
U.S. Army Research Institute
for the Behavioral and Social Sciences

The retention of quality Army enlisted personnel not only represents cost-savings in terms of training dollars and recruitment efforts but also is essential to readiness and performance. In contrast to previous years when the active force was primarily single, a steady rise in the number of married soldiers is occurring. Given the current force make-up retention efforts (e.g. policies and incentives) must take into account spouse related factors for organizational efforts to be truly effective.

The present research examines some of the spouse-related factors which soldiers have identified as most salient to reenlistment. The purpose of this paper is to report the development of a model that explains the interrelationships of these factors as they relate to soldier commitment and reenlistment.

Method

Sample

The sample consisted of 453 currently married soldiers who met the following conditions: eligible for reenlistment, within 8 months of Expiration of Term of Service (ETS) and indicating a definite reenlistment intention. Only soldiers for whom complete data were available were included in the sample.

Procedure

Soldiers were asked to answer questions from the Reenlistment Incentives and Career Decision Making Questionnaire (RICOQ) which was designed to tap a number of factors that influence reenlistment decisions. These factors include needs and their fulfillment, satisfaction with the Army, organizational commitment, occupation stress, perceptions of civilian alternatives, opinions on reenlistment policies and procedures, demographic and family variables. One section of the questionnaire was devoted to assessing soldiers' perceptions of the spouses' satisfaction with Army support services, opinion about reenlistment, the likelihood for spouse employment in a good job if the soldier left the Army and the soldier's satisfaction with his family's ability to cope with the Army way of life. In addition, information was obtained on the employment status, age and educational level of the spouse.

¹The views expressed in this paper are solely those of the author and do not reflect the views of the Department of the Army or the U.S. Army Research Institute.

Analyses

Based on a literature review, a model of expected relationships among the variables was hypothesized. Path analysis was performed following procedures outlined by Pedhazur (1982). This consisted of a series of ordinary least-squares multiple regression analyses in which variables were regressed on those which preceded them in the model. Standardized regression coefficients, or beta weights, were used as estimates of the path coefficients. Additional analyses were performed to trim the model of variables with insignificant paths to produce a restricted or overidentified model. For purposes of trimming the model, only those path coefficients at $p = .001$ were retained. Regression analyses also determined the amount of variance in a particular variable which is explained by the specified set of variables presumed to be antecedent to it in the model. Finally, a Goodness of Fit (Q) index was calculated.

Results

The initial step in the path analysis was to regress Reenlistment Intent on all of the other variables. Table 1 presents information from this and subsequent regressions, allowing comparison of the variance (R^2) explained by the general and restricted models. As the table shows, 62% of the variance in Reenlistment Intent is accounted

TABLE 1

Comparison of Variance Explained by General and Restricted Models

Dependent Variable	<u>General Model</u>		<u>Restricted Model</u>	
	R^2	Number of Predictors	R^2	Number of Predictors
Reenlisted Intent	.62	8	.62	3
Soldier Commitment to Army	.42	7	.40	4
Soldier Satisfaction with Family's Ability to Cope with Army Way of Life	.36	5	.34	2
Likelihood of Spouse Finding a Good Job if Soldier Left the Army	.19	5	.18	2
Spouse Opinion about Reenlistment	.14	4	.11	2
Spouse Satisfaction with Army Support Services	.01	2	0	0
Spouse Employed	.06	2	.06	1

for by three variables in the trimmed model. These variables were soldiers commitment, spouse opinion about reenlistment, and likelihood of spouse finding a good job if the soldier leaves the Army. The amount of explained variance in soldier commitment dropped .02 from .42 to .40 when three variables were trimmed from the set of antecedent variables. Soldiers' satisfaction with family's ability to cope, spouse satisfaction with Army support services, spouse age and to a greater extent spouse opinion about reenlistment are variables that explain over a third of the variance in soldier commitment to the Army.

In like fashion, two variables account for a third of the variance in the soldier's satisfaction with family's ability to cope with the Army way of life. These are spouse satisfaction with Army support services and spouse opinion about reenlistment. Two variables account for 18% of the variance in the likelihood of spouse finding a good job: spouse employment and spouse education. Spouse satisfaction with Army support services and spouse age account for only 11% of the variability in spouse opinion about reenlistment.

The test for the Goodness of Fit index Q compares the unexplained variance of the general and restricted models. The maximum value for Q is 1.00. Trimming the model as indicated above resulted in an overidentified model with $Q = .89$.

Figure 1 depicts the results of the path Analysis using the regression beta weights as path coefficients. As expected, soldier commitment to the Army is the primary predictor of Reenlistment Intent [path coefficient (β) = .51]. Spouse Opinion About Reenlistment also directly impacts on Reenlistment Intent (β = .35). It also does so indirectly, however, through a stronger relationship to soldier commitment to the Army (β = .42). As expected, the likelihood of the spouse finding a good job if the soldier leaves the Army, impacts on Reenlistment Intent negatively (β = -.10). Similarly, soldier's satisfaction with the family's ability to cope with the Army way of life affects soldier commitment to the Army (β = .15). Spouse satisfaction with Army support services influences soldier commitment directly (β = .14) and indirectly through soldiers satisfaction with family's ability to cope (β = .49) and spouse opinion about reenlistment (β = .25). The path from spouse opinion about reenlistment to soldier's satisfaction with family's ability to cope with the Army way of life (β = .21) is as expected.

Consistent with turnover research in both the civilian and military sectors, a path between spouse age and soldier commitment (β = .19), and spouse age and spouse opinion about reenlistment (β = .22) are significant. It is logical to assume that spouse age is highly correlated with years of direct or indirect organizational affiliation. Thus what is observed is a tenure-commitment relationship.

On the fifth level of the model which contains the employment status of the spouse a significant relationship to likelihood of spouse finding a good job if the soldier leaves the Army occurs

CONTRIBUTIONS OF SPOUSE RELATED FACTORS AFFECTING REENLISTMENT FOR ENLISTED PERSONNEL

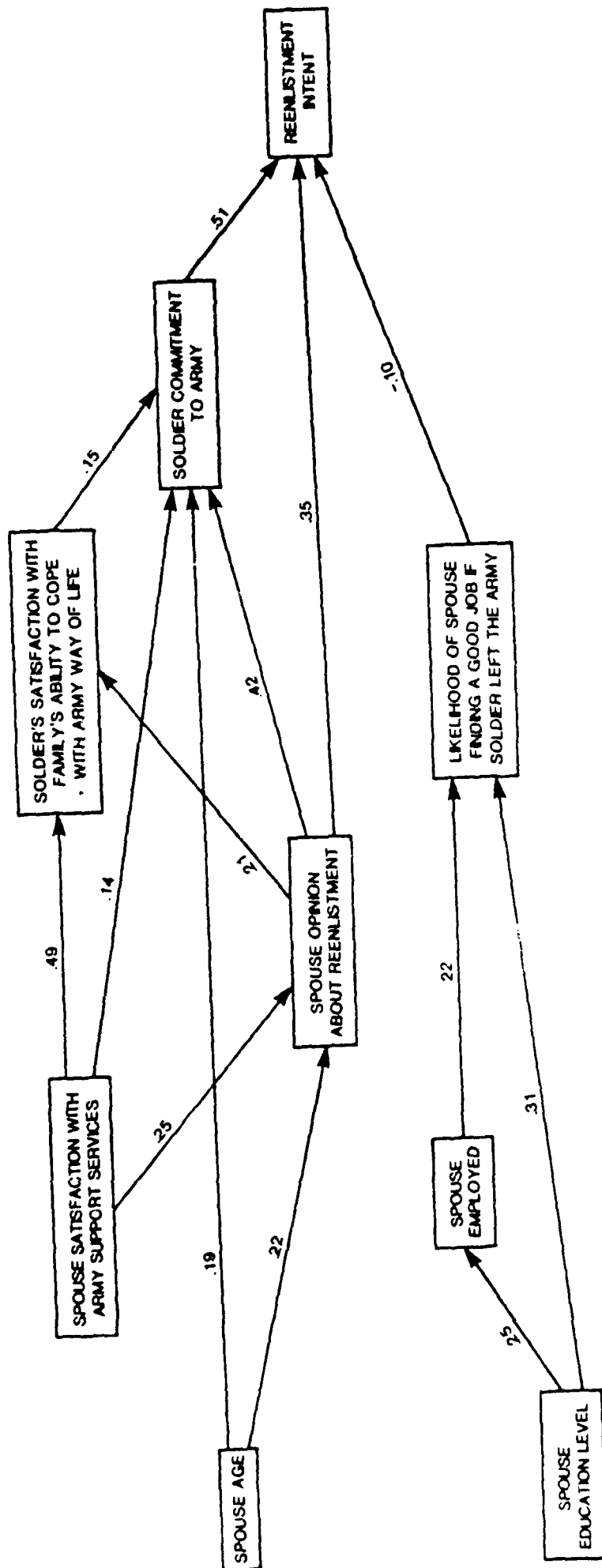


Figure 1

($P < .22$). Also, a direct path from spouse education to likelihood of spouse finding a good job if the soldier leaves the Army ($P < .31$).

Discussion

The results of the path analysis indicate that the data fit the hypothesized model well. The restricted model was able to account for nearly two thirds of the variance in reenlistment intent.

Consistent with previous findings (Smith, 1987; Peterson, 1987), soldiers' indications of their spouses' desire for them to leave or stay in the Army are highly related to soldier's intentions about reenlistment and other precursors of that decision.

The lack of paths between spouse employment and spouse opinion about reenlistment and likelihood of spouse finding a good job if the soldier leaves the Army has provocative implications. The most obvious is that employment variables appear to be operating independently of salient psychological variables that affect reenlistment intent. In other words, in this sample the employment status of the spouse does not affect the spouse's opinion about reenlistment, a key program intervention variable.

Although this research failed to demonstrate linkages between spouse employment and spouse opinion about reenlistment, a major program implication did emerge. That is, employment opportunities by themselves may not be enough for spouses of enlisted personnel, especially where the wife is highly educated. Instead employment opportunities offered by the Army and surrounding installation communities may have to become more competitive in wage and other career incentives to offset perceptions of more attractive inducements in other civilian job markets.

The present analyses has served to build a framework for enhancing our understanding of spouse-related factors affecting reenlistment. There are still unknowns. For instance, more research needs to be conducted to determine the antecedents of spouse satisfaction with support services and spouse opinion about reenlistment. In the model presented, these are key variables in designing program interventions. Perhaps spouse opinion about reenlistment is influenced greatly by the quality and happiness of the marriage unit. If so, this would not be in the foreseeable realm of Army policy intervention.

Finally, it should be mentioned that this model may be unique to the sample. This sample consisted of NCOs and a some E-4 promotables many of whom have made one or more reenlistment decisions to continue military service. Subsequently, the resulting paths in this research may not be the same for younger married soldiers facing their first reenlistment decision.

References

Pedhazur, E. J. (1982). Multiple regression in behavioral research. New York: CBS College Publishing.

Peterson, M. (1987). Army families: You're important. Soldier Support Journal. January - March, 47-51.

Smith, A. L. (1987). Aspects of Spousal Satisfaction with the Army: Soldiers' Career Decisions. Paper presented at the annual meeting of the Military Testing Association, Ottawa, Canada.

Title: Family Demographic Cohorts: Makers/Breakers of COHORT Unit Cohesion and Readiness

Authors: Joel M. Teitelbaum, Ph.D. (Department of Military Psychiatry, WRAIR, Wash., D.C.) & LTC T. Paul Furukawa, Ph.D., (U.S. Army Community and Family Support Center, Alexandria, VA.)

Presented to the Military Testing Association Conference, ManPower Trends Section, Weds., Nov. 30, 1988, Arlington, VA.

Introduction

The 1984 White Paper on the Light Infantry Concept presented our Chief of Staff of the Army's vision of a rapidly deployable lightly armed group of foot soldiers organized for maximum small unit autonomy. High performance Light units were to emerge from intensive training in cohesive groups led by highly qualified leaders. The Unit Manning System's COHORT (Cohesion, Operational Readiness, and Training) personnel movement system would provide stabilized first term troops. Vertical and horizontal bonding from the small unit to the company and battalion level would multiply fighting capacity, i.e. "soldier power".

The first Light Infantry COHORT battalion was assembled at Fort Ord, California in 1985. Full division strength was set at 10,000 plus soldiers. Down-sized combat arms units consisting of small units dispersed in the field were supplied by austere manned combat support and service support units. Each combat arms battalion numbered approximately 600 men, structured into three line rifle companies of approximately 120 men, and a slightly larger headquarters company for the entire battalion. A line company contained three platoons of some thirty men each. Three rifle squads of nine men each composed a platoon; small units had a high ratio of leaders to led; COHORT first termers made up two-thirds of troop strength within a company. Unit leaders were to bond together via a Light Leader's course, and develop light infantry culture in companies and battalions.

Soldiers' families are viewed as contributors to Light unit readiness rather than detractors from it. Hence, unit leaders were expected to provide special family supports to families such as welcoming and family support groups to mitigate the stressors of high pressure military training and deployments. Installation leadership would provide incentives such as accessible housing and community services to assist in family member adjustment and adaptation to Army life and deployment separations. Family members were, in turn, expected to bond with the unit and its mission, and to dwell at the post in a stable, supportive community during the three year COHORT unit life cycle.

However, Light Infantry unit and command leaders were unfamiliar with family issues in this new type of unit, and no systematic information or training on family leadership was available during the establishment of the prototype Light Infantry Division at Fort Ord. These shortfalls would prove to

exacerbate declines in unit cohesion and readiness as the numbers of family members per unit increased across a COHORT life cycle.

Field Research Methods and Process

The Walter Reed Army Institute of Research (WRAIR) field study of the first implementation of the Light Infantry concept began in mid-1985 shortly after the arrival of the first two COHORT battalions. An interdisciplinary team of WRAIR behavioral scientists visited Fort Ord over the next two and one-half years in an ODCSPER sponsored Human Dimensions Evaluation of the Light Infantry Concept. The study emphasized medical-psychiatric aspects of military unit and life stressors on soldier and family well being, and the effects of family stress responses on soldier combat readiness and unit group cohesion.

WRAIR's longitudinal surveys provide a cluster sample of Light Infantry troops across unit ranks. It reveals variations in individual attitudes, changes over time, and simultaneous demographic (marriage and family formation) processes during the three year unit life cycle. Two family health surveys of spouses of soldiers across the ranks in sample units were performed by mailout questionnaires. The first occurred in the second year of the unit life cycle and the second iteration was administered at the beginning of the third year.

Soldier, Family, and Unit Readiness

1) Vertical Cohesion: WRAIR's Unit Manning System (UMS) survey measured achievements of the Light Infantry Concept underway during the first year of training. However, the survey showed that unit cohesion and readiness declined drastically by the second year, as measured by trust in unit leaders and perceived individual state of readiness. Further decrements happened during the third year after replacement packages were inserted into existing COHORT units. Key attitudinal measures at the small unit level (squad and platoon) demonstrated a sharp decline of COHORT soldier confidence in small unit leadership, especially toward NCOs, between the first and second surveys. Measures of company vertical cohesion and readiness measures continued to worsen in subsequent surveys.

First term soldier confidence in their unit officers was sustained at higher levels between the first and second survey iterations. But this measure fell sharply in the third iteration and sank even lower during the fourth survey. Company level identification by first termers rose between the first to the second iteration, but then declined to a lower level by the third survey, dropping to its lowest point in the fourth survey after the addition of replacement troops. Comparison data showed that Light Infantry vertical cohesion began at levels similar to elite Army units and declined to that of conventional nonCOHORT units.

2) Horizontal bonding: Among small unit COHORT soldiers horizontal cohesion also declined somewhat over the course of the survey period. However, COHORT peer bonding was sustained better

than vertical cohesion over the unit life cycle. Horizontal cohesion dropped significantly over time only for married soldiers with families at post. It was sustained best among single soldiers living in company barracks.

3) Well Being: Soldier well being was highest among those of higher rank and greater seniority. On average, first term soldier perceptions of well being actually increased somewhat during the unit life cycle. The low was during the first iteration and the high during the third survey. Soldier perceptions of self-esteem in relation to their jobs did not decline after the second iteration and remained relatively stable thereafter. A separate survey measured well being among spouses of unit soldiers with two iterations. As with their husbands, wives' well being varied directly with rank. First term wives had the lowest level of well being from the beginning to the end of a unit life cycle. However, all unit wives perceived a clear decline in their well being between the survey administrations.

Analytic Results: Soldier-unit Demographics

Statistical analyses show that background demographic factors do not account for the variance in baseline attitudinal responses or for perceived changes over unit time among soldiers and their spouses. This study and Army records reveal that the ethnic composition of first terms in COHORT units changed little during a unit life cycle. White soldiers were slightly over three-fifths of each battalion. Black soldiers averaged one-fourth of total, and other ethnic groups rose slightly from one-tenth to one-eighth of all surveyed soldiers for over two years. Ethnic factors were not associated with the overall declines in unit cohesion and perceived readiness during a unit life cycle.

Age changes among soldiers progressed with the time gradient of unit life cycle. On average, first term soldiers were under 20.6 years of age during the first survey iteration; their mean age reached 21.3 years by the second iteration; it was 21.8 years in the third iteration. Well being rose with soldier and spouse age, but increasing age was not causally related to declining perceptions of cohesion and readiness over unit time.

Soldier rank also increased rapidly during each unit's life cycle as junior enlisted were promoted, and some were chosen to fill vacant junior NCO roles. Three-quarters of COHORT soldiers were at the E2 paygrade during the first iteration; approximately three-fifths were E3s during the second and third iterations, with E4s approaching one-third of the remaining men. By the fourth iteration, nearly two-thirds were E4s and additional E1s and E2s had been added. Seniority and promotions were positively associated with soldier perceptions of individual well being. However they were not causally related to declining COHORT soldier perceptions of unit cohesion and readiness. In short, first term enlisted promotions did not make or break COHORT unit cohesion or unit readiness for combat.

Families of NCO small unit leaders were larger, on average, than first term families, due to their longer marital status and prior childbirths. At the start-up of a COHORT unit NCO wives and their children formed a clear majority of all family members, and there were more soldiers than family members. Toward the end of a company's life cycle, junior enlisted wives and children outnumbered those of NCOs. The number of junior enlisted wives became more than half the spouses in a unit and there were twice as many family members as soldiers.

Family Issues Among COHORT Soldiers

Quality of life needs increased exponentially as family member numbers rose. WRAIR data show that most new families adjusted to their household conditions within the first year of marriage. However, COHORT unit and division leaders did not build unit training schedules or support capacity to incorporate the growing needs of new wives and young children of their soldiers into unit operations. Unit leadership efforts on behalf of families flagged over time. Extra efforts were made to welcome small numbers of newly arrived COHORT spouses and a few children during the establishment of a new COHORT unit, and when a replacement package of COHORT soldiers arrived. Family Support Groups were organized among spouses in each company and battalion on a volunteer basis to assist newcomers and offer support during deployments. A majority of wives who 'trickled in' individually felt psycho-socially isolated from their husbands' unit leaders and other unit wives, as well as alienated from the surrounding community. They did not adjust as quickly as grouped arrivals.

WRAIR surveys and interviews with soldiers and family members indicate that the fast-paced demands of Light Infantry unit training and mission schedules were the major factors depressing soldier cohesion within all units. Vertical cohesion, (trust) toward unit NCOs and officers, was the main casualty of military stressors as unit life cycles progressed. The major components of distress were unpredictability of garrison duty hours, frequency of off-post training exercises and deployments, and loss of controllable family time. Uncertainty surrounding off-duty weekends, constraints on annual leave and unpredictable daily and monthly duty hours made soldiers unavailable to their families for leisure activities. Personnel turbulence among small unit leaders and lengthy duty periods on tasks judged non-mission essential by most soldiers resulted in decreased trust in unit leadership, magnifying perceived loss of vertical cohesion.

First term wives' disaffection with unit leadership and alienation from unit support activities tended to precede decrements in cohesion and readiness among their husbands. A downward slope in unit vertical cohesion over unit time correlated with the growth rates of soldier marriage and family formation, including childbirths. Increasing negative attitudes toward an Army way of life for their families among spouses in a unit also were predictive of reduced first term reenlistment rates at the end of a COHORT soldier's first tour of duty.

A primary factor in the downward tilt of unit cohesion was soldier and spouse perception of stress. Distress about personal and family life had significant effects on unit bonding among single and married soldiers. Our surveys and interviews monitored a dramatic loss of confidence in unit leaders as military stressors accumulated over unit time. The perceived distress was strongest among married soldiers. Rising soldier marriage and family formation was associated with declining perceptions of COHORT unit leaders and combat readiness.

Family Formation and Residential Change across a Unit Life Cycle.

An exponential rise in the number of first term marriages constitutes the first COHORT family demographic revolution. WRAIR surveys, interviews and observations show that between COHORT package arrival at the duty station and the end of a unit life cycle, first term soldier marriage more than tripled, reaching one-third of a typical company's junior enlisted troops. The four survey iterations display this process over a unit life cycle as follows: Single soldiers were over ninety percent of first termers when the COHORT unit was formed. By the end of Year 2 nearly one-third of junior enlisted soldiers were married. Their proportion rose slightly during Year 3 of unit life. Company married content rose despite attrition and later replacement of attritees with packages of mostly single men.

Four-fifths of leadership cadre in each unit were married. However, rapid turnover among career NCOs and officers in these units made for frequent family relocations over the duration of unit life. Most cadre families obtained on-post housing during their first year at the duty station. In contrast, all new first term families began households off post. On-post housing of junior enlisted with families rose dramatically from one percent in the first year to fourteen percent of an average COHORT unit troops at the end of second year. Family households on or near Fort Ord were encouraged by the division as a family support.

Childbirth and Family Development

Childbirths generated a second unit-level family demographic revolution which occurred about the mid-point in the unit life cycle. The numbers of new babies rose dramatically during the second COHORT year and continued to rise in an upward sweeping curve through the end of the unit life cycle. Within each company and battalion family development traced a highly predictable expansion pattern. Typically, a small proportion of first term soldiers arrived with "ready-made families", about a dozen in each company. Many new marriages and creation of off-post households by young first termers and their even younger brides followed. Numbers of COHORT soldier children expanded from less than a dozen per company in the first year to several dozen in the later half of the unit life cycle. As one company commander belatedly noted, "Where are all these babies coming from?". As this comment suggests, most unit leaders were unprepared for two family demographic revolutions in their units.

COHORT Family Demographic Cohorts

From a social demography perspective, marriage and family growth patterns among junior enlisted COHORT soldiers differ significantly from that of their age peers in the civilian population and in other types of Army units. These first termers were more likely to be single when they joined the Army, but more of them married during their first term. About two-fifths of a Light Infantry unit's COHORT members developed families composed of young couples and early parity (i.e. childbirths)

U.S. Census figures show that the median age of marriage for all American young adults is age 25.4. Four-fifths of American young men remain single until they turn 25. Army-wide, junior enlisted soldiers are, on average, older and more married than new COHORT unit members. By the age of 22 years a third of each COHORT package had married and one-fourth had children. Most re-enlistees were married men. WRAIR surveys also showed that, in general, single soldiers perceived barracks life as unacceptable and demeaning. Many single COHORT soldiers emulated their married peers, forming households and avoiding the pressures and perceived low quality of company/battalion barracks life.

The dynamic demographic characteristics of Light Infantry COHORT soldier wives contrast sharply with their peers in the wider society. The mean age of first termer wives at the establishment of a COHORT unit was 20 years, more than three years younger than the median age for married women in American society. On average, spousal age rose but did not keep pace with soldier aging across a unit's life cycle; many younger brides continued to join the pool as average soldier age rose. These new Army wives typically gave birth to a first child over five years earlier than do most married American women. Few wives of COHORT soldiers sought abortions, compared to much higher rates among their age-peers in the civilian population. Pre-marital pregnancy was indicated in one-fourth of junior enlisted soldier marriages. Age-specific fertility rates of these young infantry wives were higher than for their age cohorts in civilian society. Their birth intervals, i.e., child-spacing, were also closer.

Accelerated rates of family formation are largely a function of enhanced stability within COHORT units in terms of employment and residence. Many single soldiers already had a close life attachment when joining the Army. A secure first job in a stable location, ample Army financial benefits, and rising income and benefits as Light Infantrymen rapidly gained in seniority and promotions, contributed to the intensity of nuptiality and spouse fertility rates. COHORT unit stabilization at one post for a three year period was a prime component of this family pattern.

Conclusion

The Light Infantry COHORT concept comprises a youthful American population grouping soldiers and their families in small, company-sized units. The processes of family growth dynamics were not anticipated when the Light Infantry concept was

drawn up. Army leadership was not aware of the demographic and social-psychological expectations of soldiers and their family members new to the Army, or their interactions in unit context.

Sizable family demographic cohorts of young wives with very young children emerged within units of Light Infantry COHORT soldiers. These families usually resided at or on the Army installation. An unanticipated spillover effect of COHORT unit personnel stabilization is that most of these new young families passed through the early stages of family formation more or less simultaneously. Wives and young children formed new demographic groupings within each unit, profoundly altering the age-sex structure of COHORT companies and battalions over time.

Unit chains of command and the Light Infantry division leadership were unable to foster an operational social climate that bonded spouses to their husbands' units or attracted them to an Army way of life. Initially high unit vertical cohesion dropped markedly as soldiers become attached to their growing families, and alienated from their unit leaders and missions.

Conflicts between family needs and unit demands escalated for soldier and spouse over unit time as the company training schedule and mission demands cumulated and intensified as part of the development of the Light Infantry division. Uncontrollable military stressors penetrated the boundaries of family privacy and exacerbated new family strains. Cumulative stress reduced soldier-unit bonding and detracted from unit cohesion and readiness. Family stress substantially reduced retention of married first termers at the end of their duty tour.

Implications

Army designs for the development of "soldier power" by means of group dynamics and training of light infantry units must take into account two fundamental demographic shifts in COHORT units:

- 1) soldier change from single status to marital status, and,
- 2) childbirths/family development among COHORT unit troops.

Leadership attention to COHORT family demographic cohorts becomes operationally crucial to maintenance of vertical cohesion within units. As families emerge as major constituents of unit organizational life, family issues develop as predictable makers or breakers of Light Infantry COHORT unit cohesion and readiness.

Rapid family formation and development can also be expected in conventional Army COHORT units as a function of the family demographic cohort effect analysed above. Predictable family demographic processes can be expected to characterize units which rely on an individual personnel movement system. Army rotation policies favoring rapid turnover of small unit leaders also cause family stress, and have adverse effects on COHORT and nonCOHORT unit cohesion and readiness. Further scientific study is needed on the dynamics of Army-family relationships in unit contexts.

GETTING BETTER RESPONSES FROM SPOUSES

Morris Peterson, Ph.D., and Susan Kerner-Hoeg,
Army Personnel Survey Division,
U.S. Army Soldier Support Center-National Capital Region,
and Emily Cato,
U.S. Army Community and Family Support Center

PURPOSE

The purpose of this paper is to describe the procedures used to obtain a high response rate from the civilian spouses of active duty U.S. Army soldiers. Special survey administration procedures were necessary because the questionnaires were to be mailed directly to the spouses and because the Army had not previously conducted a worldwide mail survey of the spouses.

BACKGROUND AND OBJECTIVES

The U.S. Army conducted in 1987 its first Army-wide survey of the spouses of active duty soldiers. The survey was unique because questionnaires and other survey materials were mailed directly to the civilian spouses of soldiers in June 1987. The results of the Annual Survey of Army Families (ASAF) are being used by Army leaders for policy and program planning. The information will help identify family programs which support and develop strong soldiers and families. The results also will provide direction for targeting family initiatives which are affordable and which encourage commitment by soldiers and their families to the Army.

The objectives of ASAF are to:

- (a) Obtain information directly from the spouses.
- (b) Assess the status of selected Army Family Action Plan issues to identify progress in resolving identified issues.
- (c) Assess family member attitudes about the Army way of life as it affects families and their perceptions of the quality of life for Army families.
- (d) Identify new and emerging concerns of families.
- (e) Supplement other evaluation and research efforts designed to improve Army support of families.

GETTING BETTER RESPONSES FROM SPOUSES - 2

The U.S. Army Community and Family Support Center (CFSC) and the U.S. Army Soldier Support Center-National Capital Region jointly conducted the survey. A second general purpose survey of Army families is planned for 1990. CFSC also is developing the Army Computer-assisted Telephone Interview (ACATI) system for quick-turnaround surveys on family issues. Other Army family surveys underway are the Army Family Research Program survey, which will be conducted during February through July 1989, and the survey of spouses of Army Reserve and National Guard soldiers, tentatively scheduled for early 1991. In 1985, the Department of Defense conducted a survey of military personnel, including active duty Army soldiers and their spouses.

SURVEY DESIGN

A stratified, systematic sample was drawn of the spouses of 20,272 active duty soldiers (8,141 commissioned and warrant officers and 12,131 enlisted personnel). The samples were stratified by rank, and oversampling was conducted to increase subsamples sizes for selected strata. The overall response rate was 62 percent (12,525 completed questionnaire were returned and 12,105 were usable). In addition, 1,755 were returned as "undeliverable," 17 were "refused," and 1 potential respondent was deceased.

The sampling procedures were designed to identify the spouses of soldiers who would be in the Army at least through 30 June 1987 and who would be available to receive the questionnaire. In addition, the sample was restricted to the "traditional" soldier-civilian spouse family. Dual military couples, sole parent families, and spouses of soldiers who have been on active duty for fewer than six months were not be included in the sample. The characteristics, needs, and concerns of these families are believed to be very different from the more traditional soldier/civilian spouse Army families. The characteristics and needs of these groups would best be examined in separate surveys.

A self-contained, optical scan readable questionnaire was used. The survey materials were mailed to the spouses of the soldiers on 15 June 1987. A tracking system permitted the second mailing of the questionnaire to be sent to those who had not returned completed questionnaires. Follow-up reminder letters were sent after each mailing of the questionnaires.

GETTING BETTER RESPONSES FROM SPOUSES - 3

ADDRESSES OF SPOUSES

The first major obstacle of conducting ASAF was obtaining the home mailing addresses of soldiers' spouses. Approximately, 5 percent of officers' spouses and 17 percent of enlisted members' spouses are not located at the same place where the soldier is assigned. There is no single, central, accurate data base for obtaining the addresses of the civilian spouses of soldiers.

The home mailing address of officers is available on the Officer Master File; however, there is no comparable information for enlisted personnel on the Enlisted Master File. By-name, by-unit rosters were developed for enlisted personnel and sent to Personnel Service Centers worldwide. Records clerks at the Centers used the Emergency Data Card (DD Form 93) to obtain the most recent addresses of civilian spouses of enlisted soldiers.

The Defense Enrollment Eligibility Recording System (DEERS) also was used. Although DEERS did not have current addresses for the majority of spouses selected for the survey, it does have spouses' first names. Thus, DEERS was used to provide the first names of the spouses of officers and for spouses of enlisted personnel whose names were not available on the DD Form 93.

PROCEDURES TO GET A HIGH RESPONSE RATE

The second major obstacle was to ensure that--after the survey questionnaires had been delivered--spouses would complete and return them. Nine major categories of procedures/activities were undertaken to obtain a high response rate. Listed below are the categories and specific procedures used.

Demonstrate Command Support

- o Send a message from the Army Deputy Chief of Staff for Personnel (DCSPER) to commanders about ASAF and other family research.
- o Send a message to Personnel Service Center (PSC) commanders, tasking PSCs to obtain home addresses for spouses of active duty enlisted soldiers.
- o Send letters from the Chief of Staff of the Army and the Sergeant Major of the Army to inform soldiers that their spouses had been selected to participate in the survey.
- o Send a cover letter to respondents from the DCSPER, explaining the purposes of the survey and encouraging cooperation.

GETTING BETTER RESPONSES FROM SPOUSES - 4

Encourage Soldier Support

- o Send a letter from the Army Chief of Staff to officers and from the Sergeant Major of the Army to enlisted personnel (whose spouses were selected to participate in the survey), encouraging their support of participation by their spouses.

Conduct Publicity Campaign

- o Conduct an Army-wide publicity campaign about the survey, coordinated by the CFSC Public Affairs Office.
- o Use Army Times for national/worldwide publicity and Stars and Stripes for additional overseas publicity.
- o Use the Army news service (ARNEWS) for publicity through installation newspapers.

Reduce Perception of "Threat"

- o Mail the questionnaires and follow-up materials directly to spouses.
- o Place address labels on a tear-off cover page so that respondents cannot be easily identified.
- o Use a preprinted code number on the back page to track returns unobtrusively.
- o Provide prepaid, self-addressed Business Reply Mail envelopes (no return address is required) for spouses to mail completed questionnaires directly to the ASAF processing center.

Personalize the Survey

- o Send the questionnaires directly to the spouses.
- o Use the first names of the spouses in the addresses.
- o Obtain the most accurate home addresses of the spouses.
- o Use the "best available" address (following pre-established procedures).
- o Use First Class Mail for the initial mailing and follow-up letter.
- o Use pictures on the cover of the survey booklet depicting Army families.

GETTING BETTER RESPONSES FROM SPOUSES - 5

Facilitate Easy Response and Return

- o Provide a self-contained (optical scan) questionnaire booklet with instructions, questions/response categories, and a specific, nearby location to record the answer(s) for each question.
- o Prepare instructions, directions, and questions/responses in an easy-to-read style.
- o Use a format for the questionnaire which facilitates easy reading and following of instructions.
- o Use Business Reply Mail self-addressed envelopes.

Do Follow-up Reminders

- o Track questionnaire returns to identify bad addresses, completed returns, and those who have not yet returned a completed questionnaire.
- o One week after the initial mailing, send "personal" follow-up letters from spouses of the Army Chief of Staff and the Sergeant Major of the Army as the first follow-up contact.
- o Six weeks after the initial mailing, send a second copy of the questionnaire to spouses with "good" addresses who have not returned a completed survey.
- o Seven weeks after the initial mailing, send a final follow-up reminder to those who received a second copy of the questionnaire.

Make It Easy for the Spouses to Contact Us

- o Include the mailing address of the sponsoring agency and the name of the person to contact about the survey.
- o Provide a toll-free telephone "hotline" for persons who want to inquire about the survey or procedures for completion and return.

Obtain Comments from Spouses

- o Encourage spouses to write their own personal observations on a "Comment Sheet" enclosed with the questionnaire.
- o Ask only that spouses report the rank of their soldier spouses and whether they are in the U.S. or overseas.

CONCLUSIONS

The response rate of 62 percent was very respectable-- particularly for a mail survey conducted during the summer months when many Army families are moving or on vacation.

The response rates for spouses matched the pattern usually obtained for soldiers: the higher the rank within enlisted, warrant, and commissioned officer personnel, the higher the response rate. The highest response rates were obtained from spouses of soldiers at the top two enlisted and officer pay grades: 75 percent for master sergeants, first sergeants, and sergeants major (senior NCOs at the E8 and E9 pay grades), and 81 percent for lieutenant colonels and colonels (field grade officers at the O5 and O6 pay grades).

As anticipated, the lowest response rate was 27 percent for spouses of enlisted members at the lowest pay grades (E1, E2, and E3). Chain-of-command involvement may be necessary to achieve a high response rate from these spouses.

The survey procedures described in this paper can be used effectively to obtain high response rates from other population groups.

A FUNCTIONAL EVALUATION OF A MODEL ARMY
PERSONNEL ADMINISTRATION CENTER (PAC) ORGANIZATION

Raymond O. Waldkoetter, Phillip L. Vandivier, and Spencer J. Murray
U.S. Army Soldier Support Center
Fort Benjamin Harrison, Indiana 46216-5700

In February 1986, the Chief of Staff Army (CSA) issued guidance to reduce company-level administration so computers would be required only at battalion (Bn) and separate company (Co) level. Shortly after this the Vice(V) CSA directed the Soldier Support Center (SSC) to set up a prototype PAC and test all new doctrine and functions resulting from various efforts to streamline/reduce administrative functions/tasks (SSC, 1986). That PAC evaluation the SSC conducted helped in building an operating prototype Bn PAC to experiment with new concepts and make recommendations for Army-wide application.

The purpose of this study was to provide an early user test and evaluation (EUTE) of the prototype PAC organization, which was conducted under normal operational conditions in garrison at Fort Stewart, GA, and during National Training Center (NTC) exercises. Two key objectives were to: evaluate the effectiveness of the Bn administrative/personnel (S-1) Office in reducing administrative workload for the Co commander (Cdr); and, determine if Personnel Service Support (PSS) workload can be further reduced, streamlined, or eliminated in the Bn S-1 Office.

As a more comprehensive strategy the Company/Battalion Administrative System (CBAS) was established by the CSA as an effort to reduce the administration at Co level. Under the leadership of SSC, many Army organizations have joined efforts to accomplish this end. The goal is to free Co Cdrs and first sergeants (1st Sgts) for their primary mission of training soldiers in preparation for war (TRADOC, 1985). In January 1987, the VCSA established a program where the CBAS staff agencies met on a monthly basis to eliminate unnecessary administrative requirements.

In war and peace, Bns and Cos are required to perform numerous administrative functions/tasks (TRADOC, 1987). These functions/tasks are generated by many functional agencies and higher headquarters in the Army. All requirements ultimately go down to the Bn and Co level for analysis, adjudication, and input. These administrative requirements have been categorized in the following functional areas: personnel, logistics (supply), finance, legal, intelligence, medical, military police, training, publications, and Command and Control requirements. There was little effective centralized control over the Bn and Co administrative workload. Traditional approaches to this problem have met only minimal success (SSC, 1987). Now an enhanced (The) Army Combat Service Support Computer System (TACCS), TACCS software packages, new organizations, and procedures are intended to reduce administration at Bn and Co level.

The views expressed in this paper are those of the authors and do not necessarily reflect the view of the Soldier Support Center or the Department of the Army.

This Army Development and Employment Agency (ADEA)/SSC appraisal study evaluated PAC force structure, office automation hardware and software, PAC functions/tasks, and procedures, and on-going advances in the area of Co-level workload reduction (SSC, 1988). The appraisal was conducted during 18-21 January 1988 at the NTC and from 29 Feb - 4 Mar 88 by the ADEA Project Officer and SSC representatives. This appraisal report documents progress made to date in the Army of Excellence (AOE) PSS effort to streamline PAC operations and makes recommendations based on user input on how the system can be further improved. Data were collected jointly by representatives from the AOE Task Force and SSC.

METHOD

Data were gathered at the 3/7 Infantry (Mech) Bn PAC, 1st Brigade, of the 24th Infantry Division, Fort Stewart, GA, and during the NTC exercises where the Bn was tested (Messier, 1988). The Fort Stewart data were collected during individual interviews with the Bn Cdr, Bn Executive Officer (XO), the S-1, the present PAC Supervisor, and a former PAC supervisor. The PAC workers were interviewed in small groups and individually as circumstances permitted. The Co Cdrs and 1st Sgts were interviewed together at each company. Data were collected by a team which utilized a standard data collection plan with given portions of the plan used for respondents at different functional levels. The data collection plan employed the following procedures: PAC workload estimation sheets requiring respondents to estimate PAC workload requirements in peace and war for 41 PAC Functions; PAC workload logs requiring individual respondents at PAC to keep track of the kind of work and time to complete work; and typing log requiring PAC individuals to keep track of the kind of typing and time to complete work.

To evaluate the effectiveness of the Office of the Bn S-1 to reduce administrative workload for the Co Cdr, a first step was to identify the deployable and nondeployable PSS functions and associated workload required by the Bn S-1 Office. The Bn Cdr, Bn XO, Bn S-1, the present PAC Supervisor, and a former PAC supervisor and workers were surveyed as subject-matter experts (SMEs). The following reference sources were employed: preliminary draft material (Army Regulation 600-8) outlining Bn S-1 responsibilities; and, AO Task Force initiatives documented in the Standard Installation/Division Personnel System (SIDPERS) users manuals. The unit commander and staff may prescribe additional duties in local Standard Operating Procedures (SOP). The surveyed Bn used the latest version of the Tables of Organization and Equipment (TO&E)/TACCS/SIDPERS.

Respondents were asked to complete the S-1 Functions/Workload Sheet which required: ratings of each approved PAC function on a priority scale of 0(not performed) to 9(extremely important), estimations of the manyears to perform each approved PAC function under both peacetime and wartime conditions; who performs each approved PAC function; and what equipment is used to perform each function. The list of 41 approved PAC functions was developed by the AOE Personnel and Administration(P&A) Task Force and SSC. The Bn was staffed with the Adjutant General School. The PAC function,

covered the recognized personnel duties ranging from personnel accounting, personnel automation, officer/enlisted management, personnel data base management, casualty operations, typing, preparation of orders, leaves and passes, through postal operations and unit manning. The designated SMEs completed the S-1 Functions/Workload Sheet.

Since a critical indicator of workload has been typing or word processing it was decided to learn whether the Office of the Bn S-1 could accommodate the typing requirements of the Co Cdr and 1st Sgt. The PAC Supervisor, unit support clerk, Co Cdrs and 1st Sgts were queried as to whether the Office of the S-1 was able to provide all word processing support for the companies from their administrative cells. Data sources consisted of answers to the following questions, and the PAC's maintenance of a typing log over a period of several months: Is typing service provided by the PAC adequate; Is typing conducted in a timely manner as measured against the Bn SOP standard; and What proportion of typing done at the PAC is conducted for the Bn vs. Cos?

To determine if PSS workload can be reduced, streamlined, or eliminated in the Office of the Bn S-1, the Bn Cdr, Bn XO, Bn S-1, PAC Supervisor, and PAC workers were observed and surveyed. Responses were compared with requirements for Bn S-1 stipulated in DA Pam 600-8 and other regulatory documents. The PAC personnel were shown the list of 41 functions and workload requirements and asked to indicate what PAC PSS workload might be reduced, streamlined, or eliminated. Questions were asked about workload requirements duplicated in different functions, reports duplicated, different data bases having similar information, company work that should be done at the PAC, and PAC work that should be done at a higher level.

RESULTS AND DISCUSSION

To evaluate the effectiveness of the Bn S-1 Office to reduce the Co Cdr's PSS workload and determine if the workload can be reduced, streamlined, or eliminated in the S-1 Office, it was decided to examine the interrelationship among peacetime/wartime priorities and workload (manyear) estimates. Frequency distributions of the four variables (Peacetime Priorities/ Wartime Priorities and Peacetime Manyyears/ Wartime Manyyears) disclosed that only Peacetime Priorities for the PAC Functions approximated a normal distribution. For this reason, nonparametric analysis techniques were used. Results from a Kruskal-Wallis One-Way Analysis of Variance showed the five designated SMEs agreed to a much greater extent regarding assignment of Peacetime Manyyear Estimates to PAC functions ($p > .05$) than for Wartime Manyyear Estimates ($p < .01$). Kruskal-Wallis results also indicated significant lack of agreement among the five SMEs for the assignment of priorities to PAC functions during both peacetime ($p < .001$) and wartime ($p < .001$) conditions.

Chi-square results indicated the frequencies of priority assignment (Lo, Med, and H) significantly departed from chance predictions (one-third of frequencies in each priority category) for both peacetime and wartime conditions ($p < .001$). The departure from

chance was largely explained by the low frequencies (15% and 20%) of PAC functions assigned a low priority rating.

Table 1
Priority Assignment Percentages

Condition	Low	Medium	High
Peace	15%	46%	39%
War	20%	35%	45%

Spearman rank correlation results indicated moderate, significant positive relationships between PAC function peace and war priorities ($r = .46$; $p < .05$); peace and war manyear estimates ($r = .60$; $p < .05$); priorities and manyears for Peacetime only ($r = .70$; $p < .05$) and priorities and manyears for Wartime only ($r = .61$; $p < .05$).

As it is generally easier to assign peacetime manyears with some disagreement between peace and war priorities, there is an overall tendency to emphasize "Hi and Med" priorities. The significant correlation among priorities and manyears suggests the PAC was generally expending relatively more time on "Hi and Med" priority functions, than on those of "Lo" priority, even though SMEs may vary in such ratings and estimates. By knowing there are situational differences in function priorities and manyears the Bn S-1 can follow a generalized order to organize priorities and manyears more favorable to Co Cdrs. Also by emphasizing automation across functions in the Bn S-1 Office the workload can be modified to improve or eliminate transactions in some instances. Resolving priority conflicts in wartime will depend most likely on the operational situation.

Since word processing is crucial under peacetime and wartime conditions, the extent to which the Bn S-1 Office could handle Co level typing was explored. Completion times (CT) for typing done at the PAC (Cos and Bn) approximated a normal distribution allowing use of parametric statistical procedures. A total of 62% of all typing was done for six Cos with 38% for the Bn.

Results of a t-test analysis indicated no significant difference ($p > .05$) between Co and Bn CTs (time to type) and between Co and Bn turnaround times (TTs/time logged-in to out). Company average CTs (12.0 min) and TTs (10.1 hrs) were recorded with Bn CTs (13.7 min) and TTs (10.0 hrs), and with the aggregate CTs (12.7 min) and TTs (10.0 hrs). A great majority (96.1%) of typing was done by deadlines though much came to the PAC due the same day.

It seemed obvious that the Bn S-1 Office was effective in modifying the Co-level workload with word-processing service for required documents. The Bn S-1 Office was readily supported for typing and the corresponding workload control. Due to the unbiased nature of Bn work assignment the Co-level tasks were completed within acceptable times with the Bn meeting its commitments. Should priorities and workload had been too strained, word processing would

reflect greater variation and the Bn S-1 Office would not have favorably directed the workload. Basically, there was little if any conflict in trying to eliminate workload since enough time was available to seek options to improve the status of Co/Bn workloads. If cross-training were emphasized greater efficiency would result in the completion of priority tasks. Further improvement to manage PAC and Co efficiency will occur as parallel functions/tasks for finance, supply, and training are automated in a more productive work environment.

The Bn S-1 and PAC play major roles in accomplishing the P&A mission at Bn level. The S-1 is responsible for overall supervision and coordination of the Bn's P&A support. When the S-1 manages PAC functions and priorities, it appears easier to allocate manyears or time to perform functions/tasks in the peacetime environment. Allocation of manyears to wartime functions/tasks is probably more difficult because mission requirements are not as reliable as in a garrison location. Assignment of priorities in peace and war most certainly depend on the workflow demands and a need to support changing command issues. Yet there is a noticeable tendency occurring in that more priorities receive high and medium ratings than low in peace and war, even with the variance among SME judgments on the scale of priority or importance. The significant correlations between and within peace and war priorities/manyears suggest the S-1 functions follow some decision-making agreement to determine which functional priorities and manyear allocations will be required to support the given mission.

With the introduction of enhanced automation procedures, the Bn S-1 Office has been able to consolidate workload requirements effectively reducing and streamlining many transactions, thereby substituting for or eliminating ineffective work methods. The great improvement in word-processing capability has permitted the S-1 to arrange priorities and workload to support more productive Co-level management for Co Cdrs and 1st Sgts. The PSS workload can be put on a more scheduled basis freeing time for those unscheduled assignments. As the S-1 has clearly shown the skill in minimizing Bn automation and word-processing production time, he/she has gained more time to assist the Commander in performing staff duties.

The S-1 and the PAC maintain a close working relation with other Bn staff elements, subordinate units, higher headquarters, and supporting organizations. Effective interaction with members of the Bn and medical staffs, subordinate and higher headquarters, and other elements is mandatory for successful P&A operations on the battlefield. In the areas of operations, supply, and training the S-1 can serve as an advisor in automating work priorities and allocating productive time to support individual and unit efficiency. Because many function priorities of the S-1 interface with the Bn and other staffs, he/she can interact to adjust staff priorities to achieve the preferred results for the Commander. Rarely will the S-1 have to lower a priority function/task when it can be shown to support the best interests of the Commander or unit.

Possibly, one of the S-1 capabilities most overlooked is the role to prevent or reduce stress. Combat stress can destroy the Co or Bn ability to carry out mission priorities and work on the battlefield. If stress is not dealt with quickly and effectively, the effects can be lethal for a soldier and unit (FM 26-2, 1986). By recording assessment of stress factors when analyzing Troop Preparedness for the Personnel Estimate, the S-1 monitors stress and recommends actions to prevent or reduce the effects of combat stress. As the S-1 serves to reduce the administrative workload for the Co Cdr, consolidate workload management in the Bn S-1 Office, and enhance work performance with stress prevention, Co Cdrs and 1st Sgts will find more time for their primary mission.

REFERENCES

- Management of Stress in Army Operations (FM 26-2). (1986).
Washington, DC: Headquarters, Department of the Army.
- Messier, P. (1988). Memorandum for Record (Trip Report). Fort Irwin, CA: National Training Center.
- U.S. Army Soldier Support Center (SSC). (1986). Model PAC Study.
Fort Harrison, IN: Directorate of Combat Developments.
- U.S. Army Soldier Support Center (SSC). (1987). PAC Standardization Questionnaire. Fort Harrison, IN: Directorate of Evaluation and Standardization.
- U.S. Army Soldier Support Center (SSC). (1988). Evaluation of a Model Personnel Administration Center (PAC). Fort Harrison, IN: Directorate of Combat Developments.
- U.S. Army Training and Doctrine Command (TRADOC). (1985). Personnel Administration Center (PAC) Drill Book (TC 12-16). Fort Harrison, IN: Soldier Support Center.
- U.S. Army Training and Doctrine Command (TRADOC). (1987). Personnel Service Support (PSS) Mission Area Development Plan. Fort Harrison, IN: Soldier Support Center.

A JOB ANALYSIS ON THE ROLE OF A POLICE OFFICER

Sheldon H. Geller, Marijane Terry
Geller, Shedletsky & Weiss, Toronto, Canada
Fred Shaw, Peel Regional Police, Brampton, Canada

Peel Regional Police Force is responsible for the policing of a large, varied territory, which has both rural and urban sectors and an ethnically diverse population. In 1987, the Force decided to review their screening and selection procedures. They had two basic prerequisites: one, that the procedures chosen be effective and, two, that minority groups would not be discriminated against by the procedures.

The procedures in place had been developed on an intuitive basis some ten years earlier, with some attention paid to what other police forces in North America were doing. While this method of developing procedures has been used in many cases, it was deemed not to be an adequate substitute for precise investigation. Accordingly, it was decided that a job analysis on the role of a police constable would be performed.

The role of a constable was chosen as constables comprise the bulk of the Peel Regional Police's strength, and all successful candidates for employment must enter the organization at this rank. Once hired, constables receive training, mandated by the Ontario Police College, and conducted by both the College and the Peel Force itself.

Method

Normally accepted technical methods were used for the job analysis, including interviews, critical incident logs, in situ observations and a literature search and review. Twenty-four police personnel were interviewed to elicit the types of tasks performed by a police constable and the skills deemed necessary for performance. Personnel interviewed

The authors gratefully acknowledge the assistance provided by the Senior Officers and Members of the Peel Regional Police.

Correspondence concerning this paper should be addressed to Sheldon H. Geller, Ph.D.,
Geller, Shedletsky & Weiss, 39 Pleasant Blvd., Suite 300, Toronto, Ontario, Canada, M4T 1K2

ranged in rank from the Chief to fourth-class constables, and represented a cross sample of experience, gender and race. In situ observations were done with police personnel on patrol.

A further method, that of the critical incident, was employed also. An informal committee of sergeants met as a group to identify critical incidents which could be used to distinguish between poor and good performance as a police constable. Six of these incidents were selected, and sergeants on the Force kept logs for a period of two weeks and recorded their observations whenever one of the incidents occurred. Incidents selected included: death notification, domestic call, traffic ticket, accident report, arrest for shoplifting and court testimony.

All of the information from the interviews was pooled, while the information from the critical incident logs was kept separate initially until a preliminary job analysis could be completed.

The Preliminary Job Analysis

Working with the pooled information from the twenty-four interviews with police personnel, the investigators rapidly determined that three recurring phrases were used to describe the job of a police constable. These phrases were "preservation of life," "preservation of peace," and "protection of property." Almost without exception, everyone questioned about the job used these phrases at some point when interviewed.

Accordingly, these three areas were thought to be the backbone of the job analysis and an attempt was made to order the findings into three distinct categories. Such categorization, however, rapidly proved to be impossible. The various tasks which had been identified as comprising the role of a police constable refused to be fitted neatly into these categories, and no amount of manipulation or different approaches to integration of the material were successful.

At this point, the material from the critical incident logs was examined, and it was discovered that nowhere in the critical incident logs did these three phrases appear. This difference was quite startling, given the almost unanimous use of the phrases by police personnel when questioned specifically about the role of a police constable.

Various explanations for this difference were examined and discarded. While the two methods of doing a job analysis were different, in theory the results could not be so dramatically different. Different personnel were interviewed from those that completed the critical incident logs, but the people chosen to participate in the job analysis were selected randomly, except for the very senior officers, and no systematic differences could be expected to occur in results.

A re-examination of the material gathered in the interview part of the job analysis was done, and it was noted that the phrases only occurred when personnel were asked to describe the tasks of the job, and did not appear in the collected skills' inventories. In fact, the identified skills were the same in both the information gathered from interviews and recorded in the critical incident logs.

The Effect of Nomenclature on Research

The investigators who performed the job analysis were experienced in job analysis, were familiar with the complexities of it, and were accustomed to approaching their work with objectivity. They were also experienced in providing psychological and training services to police and other emergency services personnel. Yet, it was not until the preliminary job analysis failed and the critical incident logs were examined and compared with the information gathered from the interviews did the disruptive effect of nomenclature become apparent.

The investigators had listened to phrases which were a stock part of the vocabulary of the police, and indeed, are entirely familiar to the public. They had accepted these phrases at face value because they understood what was meant by them, and did not question them in any specific way. However, these phrases did not contribute to the job analysis as the "preservation of life, the "preservation of peace" and the "protection of property" while representative of the work that police constables do, are not the actual tasks which are performed.

These phrases, in fact, describe general families of tasks which are performed by police constables. The verbal shorthand of the police was accepted by the investigators, as these phrases have a very familiar ring to them and an actual meaning in a conceptual sense. However, they failed the acid test in a job analysis: was the protection of property an exclusive activity when compared to the preservation of life? Was the preservation of life an exclusive activity when compared to the preservation of peace?

The Final Job Analysis

As the nomenclature of the police has an intrinsic meaning, both to them as members of an organization, and to the public as well, it was decided to include the three phrases in the job analysis. However, the nomenclature had to be kept separate, and to achieve this goal a introductory section was added to the job analysis entitled "Objectives of the Police". The preservation of life, the Preservation of Peace and Order, and the Protection of Property became extra categories in the job analysis, and were linked to a list of associated events as identified by police personnel.

The "Preservation of Life" category included such events as civil emergencies, hostage situations and natural disasters. The second category of "Preservation of Peace and Order," included public demonstrations, youth activities and community programmes. The last category, that of "Protection of Property," included such items as the protection of public spaces, private buildings and goods.

This method of design satisfied the need of both the Police Force and ourselves to see these duties, so closely connected with police work, appear in the job analysis. It had the advantage as well of maintaining a clear separation from the actual tasks and skills involved in the job, yet allowing a complete context for the job to appear in sequential order.

Once the problem of nomenclature had been solved, the remainder of the job analysis proceeded smoothly. Grouped job activities included Patrol, Enforcement of Laws, Public Assistance and Organizational Obligations. Every task or activity identified by the police in interviews could be assigned to one of these categories, exclusive of inclusion in any other.

Patrol is a distinct category in the job analysis, and describes the key job dimension of answering calls and carrying out assignments. In this dimension, the activities involved include attendance at parade, preparation of the vehicle, patrol of assigned zone, answering calls, checking property, traffic control and other duties. While this dimension is called patrol, it actually has a more comprehensive meaning, in that it covers more than the physical act of patrolling.

The second distinct job dimension is the Enforcement of Laws, and centers on the key job responsibility of the execution of legal requirements in the role of a police constable. To this category are assigned the police activities concerned with investigations necessary to lay charges, the actual laying of charges under municipal, provincial and federal laws, arrest procedures, the preparation of any material necessary for the court system and court appearances.

Public Assistance is the third dimension, and involves the key job responsibility of interaction with the public, both in the sense of active assistance and proactive policing. This category includes such activities as mediation, referrals to community services, death notifications, advice on safety programmes and substance abuse.

The last dimension is Organizational Obligations, which covers a police constable's activities in reference to internal police force requirements. It includes such activities as obeying all lawful orders from senior officers, following the regulations of the Ontario Police

Act and Force rules and regulations, maintenance of equipment, training requirements and meeting fitness and specialized knowledge requirements.

The next step was the assortment of the identified skills associated with the four basic job dimensions into the six groups of Communication, Interpersonal, Intellectual Reasoning, Organizing Skills, Personal Attributes and Physical Requirements. Each category included a list of skills and characteristics which had been identified as necessary for the successful completion of the position of police constable. Figure 1 illustrates the job objectives, job dimensions, and skills and characteristics identified in the job analysis.

Communication Skills includes accuracy in listening, questioning, clarity, and literacy. The Interpersonal section includes compassion, empathy, persuasiveness, conflict resolution and team work. Intellectual Reasoning includes decision-making, common sense, recognition, analytical reasoning and concentration. Organizing skills features evaluation, applied procedures, attention to detail and methodicalness. Personal Attributes includes sense of humour, adaptability, self-control, maturity and initiative. Physical requirements includes visual acuity, hearing and dexterity. As the list of skills is quite lengthy, only examples are presented.

To complete the job analysis, the findings were presented to a group of officers at the Peel Regional Police Force for their review, comments and suggestions, and then any adjustments necessary were completed.

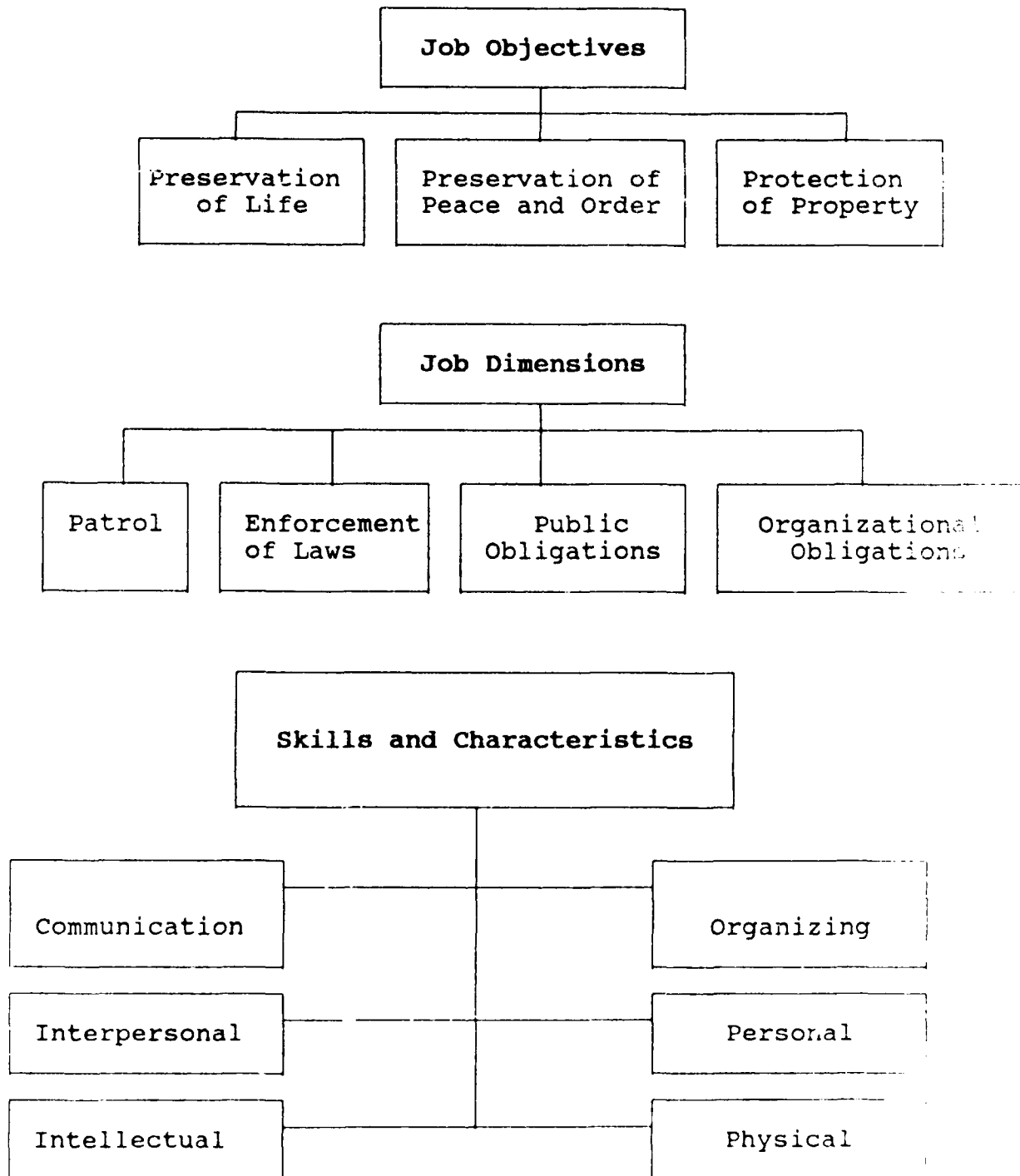
Discussion

Nomenclature, words which become a shorthand for a particular group, serves the interests and needs of that group. On occasion, it seeps into the vocabulary of others, and takes on an extra dimension beyond what it was originally intended to mean. In police work in particular, the media has helped to popularize the nomenclature of the police, as have the extensive public relations campaigns conducted by police forces.

There are many other jobs which will carry a similar type of nomenclature, and which will also resist objective analysis since the nomenclature has become part of the public domain. For example, it seems likely that positions in the military, which has a long history and a strong public awareness, will also have their own nomenclature.

From our experience, we would suggest that for other investigators attempting to perform a job analysis on a position which has a strong nomenclature, an effort should be made right from the start to isolate its effect. At the same time, a flexible form of job analysis will be needed, for the analytical method will lack acceptance without the inclusion of specialized nomenclature.

Figure 1
JOB ANALYSIS - POLICE OFFICER



The Occupational Research Data Bank:
A Key to MPT Support

1Lt Kathleen M. Longmire
Lt Col Lawrence O. Short

MPT Technology Branch
Air Force Human Resources Laboratory

A major thrust to enhance consideration of manpower, personnel, training, and safety (MPTS) factors throughout the Weapon System Acquisition Process (WSAP) has begun within each of the services. Reduced budget constraints as well as significant manpower shortages and a growing need for highly skilled personnel to operate and maintain increasingly complex weapon systems have necessitated a more efficient use of available manpower resources. Growing pressures from Congress have mandated the Air Force develop forecasts of MPTS requirements early in the design process of a new weapon system. MPTS goals and constraints need to be developed during the pre-concept and concept development stages in order to include this information in controlling documents such as the Statement of Operational Need (SON), System Operational Requirements Document (SORD), and the contract Statement of Work (SOW).

At present, MPTS considerations are not effectively included in early design tradeoff decisions or in the development of operational and maintenance concepts. We lack the tools necessary to assess the total force impact and life-cycle supportability costs of MPT decisions related to the introduction of new weapon systems or to the modification of systems currently in the inventory. The acquisition process itself does not provide detailed information sufficiently early to permit personnel and training pipelines to be properly established prior to delivery, thus causing delays in the assimilation of new systems into the operational force. While some data bases and analysis tools are available to support M, P, T, and S decisions, the Air Force is very decentralized in managing MPTS, making it difficult to integrate information from these various sources.

A major step toward MPT integration is the advent of IMPACTS, the Air Force program for Integrated Manpower, Personnel, and Comprehensive Training and Safety. IMPACTS is an acquisition management program that implements specific DOD policy regarding manpower, personnel, training, and safety and is designed to support the development of mission-capable systems that can be safely operated, maintained, supported, and contained in present and future operational environments at the lowest life cycle cost. In support of IMPACTS, research efforts are underway to define and develop a comprehensive and integrated MPTS analytic system for use in the WSAP. As part of this effort, an assessment of the tools and data bases currently available to support MPTS decisions will be made along with an in-depth analysis of their strengths and weaknesses. Examples of modeling and data systems to be examined include the Logistics Support Model (LSM), supported by data input from the Maintenance Data Collection System (MDCS); the Logistics Support Analysis (LSA) and LSA-based models available to program contractors; and the Occupational Research Data Bank (ORDB), a system developed by the Air Force Human Resources Laboratory (AFHRL) to provide researchers a variety of occupational information for use in the manpower specialties codes (MPC).

The ability to integrate information from these various data bases is very important. Techniques and software are being developed to interface these data bases; however, these efforts are still in their infancy. In the interim, MPT analysts must be able to use the information from existing data bases that is available today. Data from available sources will be especially critical during the Pre-Concept and Concept Exploration Phases in the acquisition of a new weapon system. Data collection at these two stages must include the identification of a comparable predecessor system and generation of baseline data relying on the weapon system(s) in the current inventory most similar to the emerging system. Using the existing weapon system's MPT environment as a baseline, the MPT analyst will be able to assess shortfalls in meeting new MPT requirements and be able to perform "what if" analyses on alternative weapon system designs and alternative MPT structures.

This paper will address the potential MPT applications of one existing data source, the ORDB, developed and maintained by the Manpower and Personnel Division (MC) of the AFHRL. The ORDB is an up-to-date occupational research data base containing a wide variety of both historical and current information on enlisted and officer career fields. In the following sections, we will first provide a brief description of the kinds of information available within each of the ORDB subsystems. Second, we will outline steps and provide examples which MPT analysts might currently use to establish a predecessor-system baseline data file from which estimates of MPT requirements for new and emerging weapon systems may be made. Finally, we will consider possible longer-term modifications to the ORDB system to help meet the needs of early MPT analysis.

Background/Overview

Plans for the development of the ORDB began in 1978 following the realization that, while vast quantities of information were available about Air Force occupations, the data were widely dispersed among many different organizations using many different formats and degrees of coverage. At that time, the AFHRL alone maintained 29 different kinds of computer files generated by many different sources, and other organizations (HQ USAF, AFMPC, etc.) each had their own data bases and generated numerous recurring reports, regulations, and studies. Additionally, the Laboratory housed Air Force technical reports dating back to 1943 and was the official Air Force repository of all occupational study data files generated by the USAF Occupational Measurement Center (USAFOMC). Occupational researchers in the Laboratory needed a way to consolidate this information and make it accessible to a variety of possible users.

Today's ORDB is an on-line system housed on the AFHRL UNISYS 1100/131 computer at Brooks Air Force Base, Texas. Access to the system by outside organizations is made possible via "on-line" modem-telephone lines. The design of the ORDB was created in a user friendly, tutorial environment where users are guided through the simple interface routines. The data base consists of four major subsystems from which the various types of occupational information and statistics are accessed:

- (1) Computer Assisted Reference Locator (CARL). This subsystem provides listings of occupational studies, technical reports, and other documents related to Air Force jobs. Each reference in CARL includes such information

as author/OPR, title of the reference, type of reference, a brief narrative description, and an associated list of key terms for each. Key terms input by the user are the basis for searching and accessing routines which identify and display the desired references.

(2) Aptitude Requirements Component (ARC). This subsystem contains up-to-date information from Air Force Regulation (AFR) 39-1, Airman Classification, which establishes the occupational structure of the enlisted force. The ARC contains AFSC descriptions (for ladder and career field), progression ladders, aptitude requirements, and specialty prerequisites from 1978 to the present. It also has an AFSC number change history file which tracks all changes from March 1965 through the present.

(3) Statistical Variable Subsystem. This subsystem contains statistical information on the enlisted force by AFSC, population group, and year for a total of 125 different variables. AFSC-specific data are collapsed across individual records from a number of AFHRL data files including the Airman Gain/Loss (ACL), the Pipeline Management System (PMS) and the Uniform Airman Record (UAR).

(4) CODAP Reports Subsystem. This subsystem contains selected reports from Air Force occupational studies which have been conducted by the USAFOMC. Six basic types of information are available in these computer-generated reports: tasks comprising the AFS, percentage of incumbents performing each task, relative percentage of time spent on each task, relative difficulty of each task, relative training emphasis recommended for each task, and summaries of background information such as equipment used or maintained, test or special equipment, and job satisfaction information.

Two additional ORDB subsystems use data from the main four subsystems. The Custom Reports Subsystem presents several unique custom report options to include CODAP-Statistics interface output or large volume statistical retrieval. The ORDB-SFSS Interface Subsystem allows four different SFSS procedures to be applied to data from the Statistical Variable Subsystem.

ORDB MPT Applications

As an example, data available within the ORDB may prove useful in the development of the IMPACTS Program Plan (IPP), especially in the Pre-concept and Concept Exploration Phases which often lack concrete data specific to the new system. The primary objectives of the IPP are to (1) establish MPTS goals and constraints; (2) influence the design process through participation in MPTS trade-off analyses; and (3) ensure the necessary manpower authorizations as well as qualified and trained personnel are available to support the weapon system when fielded.

To be able to use the information within the ORDB for this purpose, the MPT analyst must first identify the weapon system or systems currently in the inventory which is/are most similar to the emerging system. For example, planners within the Advanced Tactical Fighter (ATF) SEC have identified both the F-15 and F-16 tactical aircraft as the systems most comparable to the proposed ATF. As a result, the initial ATF data base will be based on information available from these two predecessor systems. Because information within the ORDB is presented by AFSC rather than by weapon system, the analyst

must next identify those Air Force specialties which work that system. In some cases, this identification is relatively straightforward and can be obtained directly from the Airmar Classification Structure Chart (AFVA 39-1). For example, AFSC 431X1, Tactical Aircraft Maintenance, is subdivided by alphabetical "shredouts" to identify specialization in a specific weapon system.

In most instances, however, a direct linkage between AFSC and weapon system is not available. For example, individuals holding the 426X2, Jet Engine Mechanic AFSC, may work on a number of different aircraft. No identifier within the code itself specifies which weapon systems are maintained by that AFSC. How then can an MPT analyst generate a list of pertinent AFSCs? One approach is to gather information from Subject Matter Experts (SMEs) who are knowledgeable on the predecessor system. Another potential source of information is the Special Experience Identifiers (SEIs) which identify special experience and training not otherwise reflected in the classification system. AFR 39-1 contains a listing of all Special Experience Identifiers along with their authorized AFSCs. Using the look-up chart available in the index of the regulation, all AFSCs authorized for an SEI may be identified. Using these AFSCs as inputs to the ORDB, the MPT analyst can begin to extract data and build a predecessor system data file.

The first ORDB subsystem to be accessed by the analyst should be the ARC. The ARC provides a narrative description of the ladder AFSC (i.e., disregards skill level--431X1E, 426X2, 423X4, etc.); a narrative description of the career field (i.e., a career area--43XXX, 42XXX, etc.); a listing of the AFSC prerequisites including minimum aptitude scores, physical profiles, whether or not women are authorized, physical work capacity, certification or license requirements, mandatory training courses, etc.; AFSC Progression Ladders as they appear in AFR 39-1; and a complete AFSC number change history with additions, deletions, and skill level changes available from 1965. The ARC is a good starting point in obtaining a qualitative description of any AFSC. It essentially automates information as it appears in the latest publication of AFR 39-1.

Having obtained qualitative descriptions of the AFSC, the analyst would next access data available within the Statistical Variables Subsystem. Frequency distributions, means, and standard deviations are generated for 125 variables and are summarized by Duty AFSC, with further breakdowns by calendar year and population group. The data are summarized at three levels: (1) career field (i.e. 43XXX, 42XXX); (2) ladder (i.e. 431X1E, 423X4); and (3) skill level (i.e. 42632, 43151F) and are available for the 5 most recent years. Variables of interest to the MPT analyst may be grouped into four categories: First, basic demographic data are available as distributions of academic education level, skill-level, age, grade, sex, race, ethnic group, and time-in-grade. Second, aptitude profiles for AFSS can be generated through accessing distributions of AFQT and ASVAB scores, broken down by Administrative (A), Electronic (E), General (G), and Mechanical (M) scores and further by MAJCOM, race, sex, and ethnic group. Third, personnel numbers are available as assigned strength by AFS, numbers assigned by major command and base, and gains and losses to the AFS by MAJCOM, TAEMS group, and current year. Finally, forcewide information is also available for all 3-, 5-, 7-, and 9-skill levels across AFSCs or for the total Air Force population. For example, the average Administrative ASVAB score for all A-level personnel in

the Air Force is available, as are average Electronic and Mechanical scores for the 5- and 7-levels. Distributions based on the entire Air Force enlisted force are provided on a number of variables (Grade, Assigned-Strength, ASVAB-CEN, etc.).

The final and most comprehensive subsystem to be accessed within the ORDB is the CODAP Study Reports Subsystem which contains selected reports from Air Force occupational studies which have been conducted by the USAFOMC. The Air Force Occupational Survey Method (OSM) is an AFSC or personnel-oriented system and, in contrast to MDC, LSA, and LCOM, is applied to most enlisted and officer AFSCs as well as some civilian series positions. In terms of task specificity, OSM survey tasks tend to be more general than MDC, LSA, or LCOM tasks. Tasks seldom describe work performed at the component level. The principle purpose of survey tasks is to provide information useful for training and job classification.

The basis of data produced by the OSM is the USAF Job Inventory which consists of a background information section and a list of duties and tasks comprising the AFSC. The Job Inventory task list is the instrument for collecting relative ratings of time spent performing tasks, task difficulty, and task training emphasis. The background information section collects standard demographic information such as primary and duty AFSC, MAJCOM, TAFMS, etc. It also collects work-environment information, such as equipment worked on and functional job area. Data collected from this background section serves as the basis for generating individual and group job descriptions for incumbents.

A number of different types of reports are available within the CODAP Subsystem. The first type is the job description which contains ordered lists of task or duty statements together with the percent of members in a group performing the task and an average percent reflecting the time a member spends performing a task. These data are also available in group summary form. Next is the variable summary report which contains frequency distributions for specified intervals, total frequency counts, and means and standard deviations on relative background and computed variables. These data are also available in percentage rather than frequencies. A third type of report provides task-level ratings of factors including learning difficulty and training emphasis. Relative difficulty ratings, defined in terms of how long it takes a person to learn to do the task, are collected from 30 to 50 senior-level NCOs, as are training emphasis ratings, defined as importance to be taught during first-term training. Special task factor reports also show tasks (along with appropriate data) mapped under areas of the Specialty Training Standard (STS), a document produced by HQ AIC which outlines all functions within an AFSC for the purpose of training personnel in that skill.

Long-Term Modifications

Having considered current MPI-specific applications for ORDB, it is now appropriate to address longer term modifications to make ORDB more applicable to early MPI analysis. Specifically, this means making the ORDB weapon system as opposed to specialty specific. Obviously, this transformation will not completely occur until the USAFOMC begins to survey by weapon system. At this point, it will be possible to include weapon system-specific information in the background section of occupational surveys. This information, along with

weapon system-specific survey techniques currently under development by AFHRL, will allow the capacity to store data and generate summary reports by weapon system. While the RIVET Workforce configurations are providing trends in this direction, necessary delays in surveying all relevant specialties and developing needed technologies will push this type of surveying into the 1990s. Until this time, certain steps must be taken to bridge the gap between our current short-term applications and the implementation of a new approach to occupational surveys. While certainly not an exhaustive list, the following options are examples of possible ORDB modification/extensions. All will require fairly extensive research efforts.

First, it is possible to build cross reference tables specifically for ORDB data which would allow us to match specialties with a weapon system and vice versa. Starting points for such tables could come from areas such as Special Experience Identifiers and related specialties currently found in Air Force classification manuals. Utilization codes currently in the ORDB are another potential data source. It may even be necessary to tap major command resources and identify subject matter experts to help with the process. However it is accomplished, this option would provide a concrete link between weapon systems and specialties for ORDB-specific data.

Second, use of currently available task level data base matching technologies such as semantic-aided analysis will allow matching of survey task data with task data from the MDCS. MDCS data are weapon system specific so this work would allow, at least by implication, ORDB task data to be "translated" into a weapon system-specific format. This technology is currently being cross validated on an extended sample of specialties using USAFOMC analyst support. Assuming current levels of matching hold, this option provides an excellent link between survey and weapon system data.

Third, we can match information on an individual's organization of assignment (available in the raw data from which ORDB was built) with the weapon system attached to that organization. While some possibility of working with classified information exists, this option allows us to look at a given weapon system and identify the specialties associated with it. Using this option would also allow generation of personnel data by cross matching an aggregation of individuals working in a given organization and their current specialties as well as their current weapon systems and lists of equipment supporting those systems. Demographic characteristics such as time in service, aptitude levels, and education along with information such as tasks performed and amount of time spent on tasks, would be immediately applicable to an expanded job typing useful for identifying the optimum person/weapon system match.

A Closing Comment

While originally developed as a research tool, the ORDB has rapidly become an important source of on-line information to an expanding list of users in the MPT community. The examples and ideas discussed here provide important extensions to existing services. AFHRL/MO is in the process of using its own resources as well as user feedback to identify the best option(s) for ORDB modification which will provide the highest payoff for early MPTIS analysis. This effort will help bring on line an important data source for use in improving weapon system acquisition.

Developing a Total Force Occupational Specification for the Canadian Forces

LCdr Dominique D. Benoit
Mr G. Jeffrey Higgs

Directorate of Military Occupational Structures
National Defence Headquarters, Ottawa, Canada

Background

Issued in June 1987, a new Defence Policy for Canada (White Paper) introduced a Total Force concept for the Canadian Forces (CF). This White Paper states that it is "both impractical and undesirable to try to meet all of our personnel requirements through the Regular Force". Many of the tasks now performed by Regular Force personnel will be absorbed by the Reserve Force. If the Reserve Force is to support the Regular Force, the "distinction between Regular and Reserve personnel must be greatly reduced. Their responsibilities must be integrated into a Total Force concept." For instance, some combat units will have a mix of Regular and Reserve personnel, each performing specific tasks. This "total" force will form the basis for the transition to a Mobilization Special Force in the event of hostilities.

This new policy has had a direct impact on the CF Military Occupational Structure (MOS) which is controlled by the Directorate of Military Occupational Structures (DMOS). The MOS is defined as the framework in which personnel are recruited, trained, employed, promoted and paid. It is maintained by means of occupational specifications which describe the job requirements for every CF occupation. The specifications are generally developed from occupational analysis (OA) data, and are divided into three sections. The first section describes the scope of employment, provides a general job description, and indicates any special requirements for the jobs in the occupation. The second section presents ideal career patterns and the occupation progression profile, while the third section, on which this paper will concentrate, lists the specific duties, tasks, knowledge and skills required for the jobs in the occupation.

The need for the ability to compare all three components of the CF (Mobilization, Reserve and Regular), on the basis of common performance standards for essential tasks, was generated by the ongoing development of personnel policies for the Total Force concept. To adequately reflect the occupational

requirements of the three components, and to assist in the development of selection, training, employment and career development plans, the comparison must be made through related occupational specifications. Currently, however, the job requirements for the three components are described in three separate documents.

DMOS was therefore faced with the challenge of producing a document that would specify concurrent occupational requirements for the three components. Consequently, in April 88, a feasibility study was conducted to adapt the current specifications to the Total Force concept.

Current Format

Figure 1 shows the format currently used for a CF Regular Force occupation. In the left hand column, tasks are organized into duty areas. This section of the specification describes through a codified system the type of task involvement, as well as the knowledge and skill levels related to each task. For the usual eight qualification levels within the occupation, the codes define the level of involvement in the task from "assist" to "supervise", the level of knowledge required to perform that task from "basic" to "complete", and the skill level required from "limited" to "highly skilled".

OCCUPATION SPECIFICATION - TRADE SPECIFICATION - SUP TECH 911		SECTION 3 - TASK LISTING							
Serial	Task	QL3	QL4	QL5	MCPL	QLAA	QLAB	QLC	QLD
<u>OPERATIONS</u>									
1001	Demande materiel:								
	a. Internally within the CFSS; and	Db1	Db2	De3	DSec3	DSd3	Sc-	Sc-	Sc-
	b. externally from other than CFSS sources.	-	Aa1	Db2	DSb2	DSec3	DSd3	DS-	DS-

Figure 1. Current Specifications Format

In the specification, the tasks and the knowledge, skill and involvement levels are derived from OA data. Using the cluster diagram, the jobs within each qualification level are specified. This process determines the Regular Force occupational requirements. A Regular Force occupation normally comprises several different jobs performed by the members of the occupation.

Once the Regular Force specifications are developed the mobilization jobs are then identified by applying

person-one job" principle to the Regular Force specifications. However, since OA data is routinely collected in peacetime, many mobilization jobs must be identified with the assistance of subject matter experts; they develop the job descriptions that do not appear in the current peacetime picture and modify the peacetime ones to meet mobilization requirements. Each job is then organized into a mobilization sub-occupation, for which its own specification is produced.

At present, the Reserve Force specifications are developed independently from Regular and Mobilization specifications, solely by subject matter experts who identify the task, knowledge and skill requirements for each Reserve occupation. Reserve Force occupations are generally a combination of several mobilization jobs, but they are more narrow in scope than the Regular Force.

The application of these current procedures creates several problems, some of which are directly related to the Total Force concept. Mobilization plans require conversion of Regular and Reserve personnel to a mobilization occupational structure. As it stands, there is no assurance that the capabilities and limitations of Reserve personnel are congruent with the job requirements for the Regular and Mobilization forces. Regular and Reserve Force specifications are not necessarily compatible, and automatic conversion is not feasible.

The relationship between the task performed and the knowledge and skills requirements is difficult to understand because of the codified system below each qualification level (see Figure 1). With simple tasks, knowledge and skills are often inherent in the task itself and do not need to be specified. More complex tasks, however, do not lend themselves well to this process since they may require a wide range of knowledge and skills to execute them. Consequently, this codified system does not provide an accurate description of related knowledge and skills required to perform a job. Moreover, during the OA process these secondary factors are not collected as they directly relate to tasks, but only as they relate to the jobs performed at each qualification level.

Because our Regular Force is composed entirely of volunteers, the OA system provides a career progression profile, which is reflected in the specifications by eight qualification levels. OA data analysis has not been reversed over the full range of occupational employment. These levels of responsibility, Technician, Supervisor and Manager are used in the Regular Force and the Reserve Force. Mobilization forces, however, are normally the structure of work found in the military, making it impossible to determine a progression of civilian and military jobs. An

the Total Force concept evolves, it becomes essential that those who will assign personnel to positions during Mobilization be able to identify which individuals are able to do which jobs. This can only be achieved if the information for the three components is comparable, and readily available, which is not the case now.

In addition, because of the different procedures and formats used in the development of the specifications the information in the current format cannot be easily displayed and manipulated since it is not issued from a common data base.

The key to the problem and the major challenge, therefore, was to find a format that could describe the occupational requirements of the three components in a single document that could display the detailed job information concurrently.

Proposed Format

Since it was premised that the Reserve and Regular force occupations were to be developed by building from the basic mobilization jobs, the approach to gathering and utilizing OA data would require a shift in emphasis. Instead of first determining the peacetime jobs, detailed job typing would have to be refined to assist in specifying the mobilization jobs which would be supported by job descriptions derived from OA data. These job descriptions in turn would form the basis for developing the mobilization structure. The process would, however, still be supplemented by subject matter expertise to develop those mobilization jobs not identified from the peacetime data. Some of the mobilization jobs would be combined to form the Reserve force occupation, while the specifications for the Regular Force would be a composite of all of the jobs identified through the OA process for each of the four levels of occupational employment.

These major differences in our approach to analysis will have certain repercussions on the entire OA process. Because of the narrower scope of the mobilization jobs, analysts would have to be more specific when they define the jobs. In addition to refining of the job typing process, inventory of tasks might expand as the specificity level of the tasks required to define the mobilization jobs would increase. A more detailed level of specificity would also impact on knowledge and skill statements. Because they are not tied to individual tasks, skills and knowledge statements would have to be specific enough to supplement and amplify the job requirements at each of the four levels. Questionnaire administration would be affected as both Regular and Reserve force personnel would have to be surveyed to objectively determine the differences in the performance between these two components of the CF.

With all of these factors in mind, a new format was conceptually developed, and is shown at Figure 2. Duties and tasks are identified for each of the four principal levels of qualification which are comparable to the commonly found levels of employment. The new format provides explicit knowledge and skill statements reflected for each qualification level, defines the job requirement for all three components of the CF in a single document, and, for commonly performed tasks, defines it to the same standard.

OCCUPATIONAL SPECIFICATION		
SUPPLY TECHNICIAN - MOC 211		
QUALIFICATION LEVEL - APPRENTICE		
DUTIES/TASKS	MOBILIZATION MOCs	RES REG
	A B C D E F G H I J K L M N P Q	N M A C
DEMANDING		
PREPARE SUPPLY DOCUMENTS TO DEMAND MATERIEL
VERIFY PRIORITY CODES ON DEMANDS (IOR, CODE 1, ETC.)

KNOWLEDGE LEVELS		
1 = BASIC KNOWLEDGE		
2 = DETAILED KNOWLEDGE		
3 = COMPREHENSIVE KNOWLEDGE		
4 = COMPLETE KNOWLEDGE		

KNOWLEDGE REQUIRED		
MILITARY WRITING	1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	1 1 1 1
MATERIEL PRIORITY CODES (MPC)	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 2 2 2
MATERIEL DEMAND PROCEDURES	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 2 2 2

SKILL LEVELS		
1 = LIMITED SKILL		
2 = SEMI-SKILLED		
3 = SKILLED		
4 = HIGHLY SKILLED		

SKILL REQUIRED		
DETERMINING DEMAND PRIORITIES	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 2 2 2
DETERMINING ALTERNATE/SUBORDINATE ITEMS	2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2	2 2 2 2

Figure 2. Proposed Specifications Format

This format has several advantages. First, by displaying occupational requirements for all components in one document under one format, it is easily cross-referenced. Second, it

eliminates an obscure codification system and displays the knowledge and skills requirements as identified in the various jobs included at each qualification level. Third, it lends itself to automation, an essential implementation requirement: the information can be displayed in different ways to assist various users. Fourth, because the analysis starts with mobilization jobs and builds upon them, this format makes direct use of OA data, by organizing information more efficiently. Lastly, it reduces the number of iterations the job information goes through between OA and training.

Conclusion

This proposed format is the first iteration of an integrated occupational specifications. The new format, and the refinement in OA procedures are being implemented for the studies in progress. These are Combat Arms Officers, Air Navigators and Airborne Electronic Sensor Operators, Physical Education and Recreation personnel, and Communications and Electronics Technicians. They will be used to streamline the proposed format, by evaluating its strenghts and weaknesses in an applied environment.

References

Challenge and Commitment, A Defence Policy for Canada (1987). Ottawa, Canada: Government Publishing Centre, 89p.

An Application of KSA Analysis to Selection and Training Decisions

Janet George Irvin
Universal Energy Systems, Inc.
Janet H. Blunt
Tulsa, Oklahoma

Introduction

When system development results in the creation of new positions or jobs, an important issue is the identification of appropriate personnel to fill each new position. A relevant consideration is the knowledges, skills, and abilities (KSAs) that will be required in the new jobs. Because KSAs describe the qualifications of the people needed to fill the positions, they can be used as selection criteria or, if appropriate personnel are not available, they can serve as the objectives of training programs designed to produce qualified personnel.

In theory, the requisite KSAs for a new position define the ideal job candidate and the selection process mandates the evaluation of individuals from both within and outside the organization in an effort to find the best candidate. In practice, it is very often the case that the search for appropriate job candidates is limited to personnel within the organization. Particularly in large organizations, the strategy can be as gross as identifying internal job types or job classifications that are perceived to contain personnel with KSAs that approximate those required in the new position. However, the identification of appropriate job classifications from which to draw potentially qualified job candidates typically does not receive systematic consideration. The present study sought to develop a methodology for addressing this problem.

The goal of the present analysis was to identify the types of personnel best suited to fill seven positions required to support a new Strategic Air Command (SAC) semi-automated system for aircrew mission training. Because SAC had tentatively identified the Air Force Specialties (AFSSs) that would be assigned to the newly created positions, this study sought to determine the suitability of those assignments by analyzing the "core" KSAs required for the duties and tasks performed in each of the seven new positions. Specifically, KSA information associated with each of the new positions was collected, analyzed and compared to similar data associated with the AFSSs tentatively selected by SAC in an effort to determine the suitability of those AFS selections; any mismatch between existing and required KSAs would indicate a training need.

Method

The New System and the Target Positions

The Strategic Training Route Complex (STRC) is a semi-automated training system designed to provide SAC aircrews with diverse, realistic combat aircrew training. In simulated missions, aircrews will have the opportunity to realistically simulate penetration of, and weapons delivery in, hostile environments. The Route Integration Instrumentation System (RIIS) includes

the hardware, software, communication links, trained personnel, and other resources required to prepare training plans and scenarios, schedule training missions, monitor training aircraft, collect data, prepare and present briefings/debriefings, and maintain archives. At this time, eight new positions are anticipated to support the STRC/RIIS. All of these positions, with the exception of Lobby Dispatcher, were addressed in the present analysis. They are listed in Table 1 with the number of AFSs tentatively assigned to each.

Table 1. Number of AFSs Tentatively Assigned to STRC/RIIS Positions

STR Operator	2
Debriefers	7
Briefer	7
Scenario Generation	8
Planning/Scheduling	5
Activity Monitoring	6
Resource Monitoring	3

Subjects

Identifying KSAs requires the input of subject matter experts (SMEs) familiar with the tasks to be performed in, or knowledgeable of the STRC/RIIS. In this case, however, the positions of interest do not yet exist, so incumbents could not be interviewed. Acceptable alternates, however, were individuals familiar with similar systems and their associated tasks, individuals familiar with the STRC/RIIS, and individuals performing STRC/RIIS-type tasks manually.

For all seven positions, SAC personnel with these qualifications from varying AFSs served as SMEs in the data collection process. Table 2 identifies the number of SMEs who evaluated each position and the number of different AFSs they represented. In order to obtain some measure of response reliability, an effort was made to solicit input from at least three SMEs for each position. This was possible in all but two cases. At the time, no Air Force personnel were performing duties similar enough to those anticipated in the Resource Monitoring or Activity Monitoring positions to warrant interviewing them. In addition, it was felt that both positions were sufficiently complicated, yet also uncertain, to justify interviewing someone who was intimately knowledgeable of the proposed system. Therefore, only one person was consulted, a person who provided significant input to the system design and understands how these positions are meant to function.

Table 2. Subject Characteristics for Each STRC/RIIS Position Analyzed

Position	Number of Subjects	Number of AFSs Represented
STR Operator	6	2
Debriefers	5	3
Briefer	6	4
Scenario Generation	6	2
Planning/Scheduling	7	6
Activity Monitoring	1	1
Resource Monitoring	1	1

Materials/Instruments

Due to the complexity of the tasks to be evaluated and the fact that the positions do not exist, it was decided that the KSA data collection effort would be best conducted through interviews. This would allow for greater flexibility in providing task information and eliciting KSA information. Three types of interview instruments were developed for this purpose: Duty/Task Hierarchies, Validated Task Listings, and KSA Listings.

The Duty/Task Hierarchies graphically display the duties associated with the operator positions and, in some instances, the individual tasks that comprise each duty. (See Figure 1.) The Validated Task Listings provide even more detailed task information for each duty. Both the Duty/Task Hierarchies and the Validated Task Listings for the STRC/RIIS positions were developed by the system designers. The third instrument, the KSA Listings, was developed following a three-step process. First, SAC identified all the AFSs from which personnel are expected to be drawn to fill the new STRC/RIIS positions. Next, Air Force Regulations (AFRs) 36-1 and 39-1, documents that provide descriptions of knowledge, skills and abilities contained in Air Force specialities, were used to extract the knowledges and skills/abilities associated with each AFS. For example, eight AFSs have been tentatively selected for the STRC/RIIS Scenario Generation position. Therefore, the KSAs associated with these AFSs were compiled into one KSA Task Listing for the Scenario Generation position. Finally, researchers familiar with STRC/RIIS and trained in job analysis techniques developed additional KSAs that appeared to be relevant to performance in the STRC/RIIS positions. Thus, the final KSA Listings included those KSAs extracted from the AFRs 36-1 and 39-1 as well as those developed by the researchers. (See Figure 2.)

Figure 1. Duty/Task Hierarchy For The Scenario Generation Position

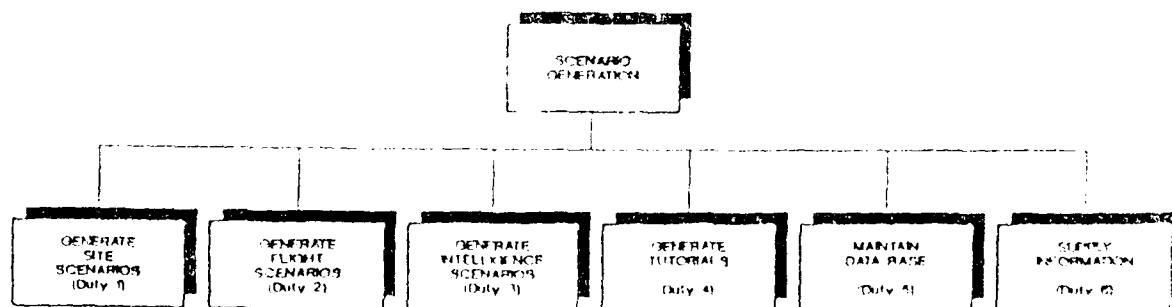


Figure 2. KSA Task Listing For Scenario Generation Position

SAMPLE KNOWLEDGES

1. KNOWLEDGE OF ELECTRONIC PRINCIPLES, INCLUDING THEORY OF TRANSISTORS AND SOLID STATE COMPONENTS AS APPLIED TO ACQUISITION AND AUTOMATIC RADAR. (30353, 30373)

___ 1.0 ___ 2.0 ___ 3.0 ___ 4.0 ___ 5.0 ___ 6.0

2. KNOWLEDGE OF LOCAL PROCEDURES OR INSTALLATION, MAINTENANCE, AND REPAIR OF AUTOMATIC TRACKING RADAR EQUIPMENT. (30353, 30373)

___ 1.0 ___ 2.0 ___ 3.0 ___ 4.0 ___ 5.0 ___ 6.0

3. KNOWLEDGE OF MAINTENANCE DATA PROCESSING PROCEDURES. (30353, 30373)

___ 1.0 ___ 2.0 ___ 3.0 ___ 4.0 ___ 5.0 ___ 6.0

SAMPLE SKILLS/ABILITIES

1. SKILL IN INTERPRETING ORDERS, BLUEPRINTS, WIRING DIAGRAMS, AND SCHEMATIC DRAWINGS. (30353, 30373)

___ 1.0 ___ 2.0 ___ 3.0 ___ 4.0 ___ 5.0 ___ 6.0

2. SKILL IN INSTALLING AND MAINTAINING AUTOMATIC TRACKING RADAR, ASSOCIATED IDENTIFICATIONS EQUIPMENT. (30353, 30373)

___ 1.0 ___ 2.0 ___ 3.0 ___ 4.0 ___ 5.0 ___ 6.0

3. SKILL IN REPAIRING AUTOMATIC TRACKING RADAR AND ASSOCIATED IDENTIFICATION EQUIPMENT. (30353, 30373)

___ 1.0 ___ 2.0 ___ 3.0 ___ 4.0 ___ 5.0 ___ 6.0

Procedure

Each interview was conducted by individuals experienced in interview procedures. After preliminary introductions and an explanation of the overall project, the interviewer explained the purpose of the interview, i.e., to solicit the subject's expert opinion on the relevancy of certain KSAs to performance of STRC/RIIS duties and tasks. The KSA Listing and the Duty/Task Hierarchy were provided and instruction given in how to use the task information for each duty to evaluate each KSA. Specifically, the subject was asked to consider one of the duties for the position and evaluate whether each KSA contributed to duty performance. Once all the KSAs had been evaluated for the first duty, the subject considered the second duty and again evaluated each KSA for relevancy. The subject continued until the complete list of KSAs has been evaluated for every duty in the STRC/RIIS position. Throughout this process, the Validated Task Listing was available if the subject required a more detailed understanding of the task requirements of each duty. In addition to the KSAs provided for consideration on the listing, the subject had the opportunity to suggest other relevant KSAs that were absent from the list.

Results

In the interest of providing the reader with relevant information regarding the methodology that was used, the results of one representative position, Scenario Generation, will be described in detail.

Personnel assigned to the Scenario Generation position are responsible for creating training scenarios, briefing materials, and training materials using a silicon graphics computer system in an interactive mode. The six primary duties anticipated for this position are: (1) generate site scenarios; (2) generate flight scenarios; (3) generate intelligence reports; (4) generate tutorials; (5) maintain database; and (6) supply information to requesting users. (See Figure 1.) In order to focus the analysis on the duties critical to position performance, only duties 1 through 4 were the subject of the previously described KSA analysis.

From AFRs 36-1 and 39-1 for the eight AFSs tentatively assigned to this position, 49 knowledges and 31 skills/abilities were extracted. In addition, four knowledges and two skills/abilities were identified by the researchers as potentially relevant to performance in the position of Scenario Generation. This list of 53 knowledges and 33 skills/abilities served as the stimulus for the six individuals who evaluated the Scenario Generation position.

Interrater reliability was estimated by calculating a Pearson product-moment correlation for all possible pairs of raters. These correlations ranged from .13 to .52 with an average value of .32. All correlations except one were significant at the .001 level.

With multiple subjects or raters, it was possible to identify those KSAs consistently cited as relevant to position performance and to weed out or eliminate those identified by very few subjects. It was determined that "core" KSAs would be defined as those agreed to by the majority of the subjects; in this case, by at least four of the six subjects. Using this criterion, a total of 23 knowledges and 16 skills were identified as "core" KSAs.

In order to evaluate the suitability of each tentatively assigned AFS, the percentage of each AFS's KSAs identified as core KSAs was calculated. For this example, two AFSs appeared unsuitable for staffing this position because their percentages were very low. This statistic was not used to make an absolute suitability decision. The specific core KSAs were evaluated individually in order to determine if, although they were few, they would contribute significantly to position performance and they were unique to that AFS. Neither was the case for these KSAs.

The remaining six AFSs contributed a much higher percentage of their KSAs to the final list of core KSAs, but again, they were evaluated on an individual basis. Some AFSs contributed few KSAs to the initial list so that even if their percentages were high, they might represent relatively few core KSAs compared to other AFSs. Again, the concern was with their significance in terms of their potential contribution to position performance.

The core KSAs contributed by the six remaining AFSs indicated that all six would contribute uniquely to position performance. No one AFS compensated for the KSAs contributed by the others. Therefore, a team approach, drawing on the expertise in each AFS, appears advisable.

In addition to evaluating AFS suitability for the position as a whole, the KSAs required for the separate duties were evaluated. Of the four duties evaluated for the Scenario Generation position, one, "generate intelligence reports," required relatively fewer KSAs than the others. While anyone possessing the KSAs required to perform the other three duties certainly could generate intelligence reports, they would be very much overqualified. If it were warranted, perhaps due to a shortage of fully qualified personnel, this duty could be reallocated to personnel possessing only the KSAs required to generate intelligence reports.

The final evaluation of the core KSAs addressed those that were not drawn from AFRs 36-1 and 39-1, i.e., those suggested by the researchers and only that resulted from subject comments. In this instance, the six KSAs added by the researchers were agreed to by all of the subjects. Five additional KSAs were suggested by at least one subject. Since these KSAs were not identified in existing descriptions of the qualifications of the personnel in the tentatively selected AFSs, they should be reviewed to determine if they represent a training requirement.

Summary and Conclusions

The goal of the present analysis was to address a very practical issue: When new positions are identified as the result of the development of a new system, is there a way to systematically evaluate those positions in advance to determine if qualified personnel are available and what, if any, training will be required? This was not a study in the conventional sense, but an attempt to try out a methodology given some very real constraints. First, the Air Force had already identified the AFSs from which they intended to draw personnel to staff the new positions. Certainly they considered the probable match between AFS qualifications and position requirements, but the choices were also influenced by existing AFSs within SAC and anticipated budget constraints. Second, the system itself was still undergoing development and changes, both to the overall system and the individual positions, could still be expected. In fact, this KSA analysis raised a number of questions that may still result in system and position modifications. Finally, although the researchers became very familiar with the new positions, their role was not to make decisions as to suitability and training requirements, but to make recommendations and to provide information summarized and organized in such a way that operations personnel could utilize it in making final decisions.

As a methodology, this KSA analysis appears to have some practical value. Useful task information was gathered as well as the KSA requirements for each position, and reasonable recommendations could be made to assist the Air Force in anticipating staffing and training requirements. Given the system design questions raised by the KSA analysis itself, the need to address staffing and training issues at an even earlier point in the design process was made clear.

AFFORDABLE AND CREDITABLE PROCEDURES FOR
DETERMINING OCCUPATIONAL LEARNING DIFFICULTY

Phillip A. Davis, SQNLDR RAAF

Air Force Human Resources Laboratory (AFHRL)
Manpower and Personnel Division
Brooks AFB, Texas

New technologies are needed to estimate Manpower, Personnel, and Training (MPT) requirements and tradeoffs during the planning and concept development stages of new and modified weapon systems. To this end, an extremely useful decision-making tool is the measure of Occupational Learning Difficulty (OLD) for each Air Force Specialty (AFS). The measure of occupational learning difficulty is used for setting appropriate aptitude standards, as stated in AFR 39-1, Airman Classification Regulation, for both entry-level and cross-specialty transfer requirements. It is also used in the Air Force person-job match algorithms for determining individual assignments to specialties, as described by Weeks (1984).

The work behind the OLD measurement began back in 1973 when the Air Force Military Personnel Center (AFMPC) requested that the Air Force Human Resources Laboratory (AFHRL) conduct research to develop an objective procedure to aid in establishing relative aptitude requirements for enlisted occupations. After 10 years of extensive research, AFHRL developed a technology for this purpose. The technology produced measures of OLD. Occupational learning difficulty is operationally defined as the time it takes to learn to perform an occupation satisfactorily (Mead & Christal, 1970).

In deriving measures of OLD, three types of occupational information were employed: (a) task time-spent ratings provided by incumbents, (b) supervisory ratings of task difficulty, and (c) benchmark ratings of task learning difficulty obtained through evaluations by contract personnel. The first two measures are available from the Air Force Occupational Measurement Center (OMC). Benchmark ratings were necessary because supervisory ratings of task difficulty only provided information concerning the relative order of tasks within occupations. Consequently, supervisory ratings were not comparable across occupations. However, benchmark ratings of task learning difficulty which are based on task-anchored benchmark rating scales (Burtch, Lipscomb, & Wissman, 1982) are comparable across occupations within a given aptitude area. Benchmark ratings were collected by contract personnel for this purpose. OLD measures were derived for more than 200 enlisted AFSs.

Following the initial data collection by contract personnel, research was undertaken (Garcia, Ruck, and Weeks, 1985) to enable the transfer of this technology to an operational setting. The procedure thus developed used Air Force personnel from OMC to routinely collect benchmark ratings. This was to provide up-to-date OLD estimates for any AFS. Teams of OMC staff would conduct interviews/task observations and then rate the tasks on the 15-point benchmark scales. This procedure proved impractical to support, due to travel and manpower cost requirements. Consequently, it was never fully implemented.

To properly transfer the learning difficulty measurement technology from a research to an operational setting, it is necessary to develop a practical (must be fast and cheap) procedure for collecting reliable task difficulty data on the 25-point benchmark scale. One potential solution is to develop a procedure involving mail surveys to collect judgmental ratings from Subject Matter Experts (SMEs). This paper describes the approach taken and results obtained in the quest for a solution using mail surveys.

The research was conducted in two phases. Phase I involved only one AFS and was primarily aimed at developing and testing the survey instrument and procedures whilst providing initial data for analysis. Phase II involved the collection and analysis of data for eight AFSs using the modified survey instrument and procedures based on experience gained in the pilot study.

PHASE I

Method

There were three criteria used in selecting the single AFS for analysis in Phase I. Firstly, original benchmark learning difficulty ratings collected by the contractor personnel had to be available for the AFS. Secondly, there should have been minimal or no change to the structure of the AFS since the original study. Finally, there should have been no significant change to the nature of tasks completed by the specialty. This would ensure that current incumbents understood the tasks that were being rated. The specialty selected was Instrumentation, Air Force Specialty Code (AFSC) 316X3. This AFSC is from the electronic aptitude area.

The mailable survey instrument consisted of five items: (a) a motivational cover letter with detailed completion instructions; (b) a background information sheet for the collection of demographic data about each rater; (c) the electronic benchmark rating scale booklet, which contained the 25-point rating scale and explanations of each task as developed by Burtch et al.; (d) the list of 60 tasks to be rated on a rating form (these 60 tasks were the same 60 originally used by the contractor); and (e) a follow-up information form which was designed to elicit useful information about the understandability of the survey they had just undertaken.

Five groups of raters were selected to complete the survey: (a) 10 SMEs, 7 and 9-skill level enlisted members randomly selected from AFSC 316X3; (b) AFHRL research behavioral scientists, all with research experience; (c) occupational analysts from OMC, all with experience in developing and analyzing occupational surveys; (d) 7 training developers from OMC, all with experience in task analysis and in the USAF training system, and (e) novices, predominantly young adults with little or no Air Force experience. Groups, except the SME group, consisted of civilian, military officer and military enlisted members. Surveys were sent to the above groups.

Analyses of the data collected in the pilot study had three primary goals. These were to (a) provide suggestions for improvement modifications to the survey instrument, (b) determine the validity of the SME's responses, and (c) compare measures of OJD generated from SME benchmark data with data generated from contractor data.

Results

The response rate for the 316X3 SMEs was 23 responses to 40 mailed questionnaires. Nonetheless, there were sufficient responses to provide a reasonable basis to meet the goals of Phase I. Response rates for the other non-SME groups were very good with almost everyone responding. Responses to the follow-up information form suggested that no significant changes to the survey package were required. Responses for each of the respondent groups were averaged to form a mean group rating. Intercorrelations between SME mean group ratings, contractor ratings, and original supervisor task difficulty ratings from the OMC OA study were calculated using the ASCII Comprehensive Occupational Data Analysis Programs (CODAP) CURVES program¹.

As can be seen in Table 1 the SME ratings correlated well with contractor ratings ($r = .75$) and even better with supervisor ratings ($r = .86$). The contractor ratings correlated at the same level as the SME ratings with the supervisor ratings ($r = .75$).

Table 1. Correlations (r) and Average Group Ratings

Group	r with Contractor Ratings	r with Supervisor Ratings	Av Rating (Across 60 Tasks)	Av Abs Dif with Contractor Ratings
Contractor	1.00	.75	13.13	0.0
316X3 SMEs	.75	.86	14.03	1.78
AFHRL Scientists	.87	.83	12.77	1.82
OMC Occup Analysts	.79	.84	12.03	2.08
OMC Eng Developers	.74	.77	12.40	1.90
Novices	.55	.45	11.68	2.63

It was important that the SME ratings fall close to the criterion level on the 25-point scale. The average SME rating across all 60 tasks was 14.03, slightly above the contractor average rating of 13.13 (see Table 1). An OLD was generated for both SME and contractor rating data. These were determined using ASCII CODAP and is the Average Task Difficulty Per Unit Time Spent (ATDPUTS) of first-term airmen multiplied by 10. This uses task difficulties established on the 25-point benchmark scale for all tasks in the AFS. These benchmarked values are extrapolated from the best fit linear relationship between the original supervisor task difficulty ratings and the 60 benchmarked tasks. The OLD using SME data was 129 and 122 using contractor data.

The four non-SME group mean ratings correlated well with the contractor and supervisor ratings (see Table 1). The average rating across all 60 tasks was in each case below the contractor rating. The average absolute difference between group mean ratings and contractor ratings ranged from 1.82 to 2.63, all worse than the SMEs at 1.78.

¹The original CODAP system that was developed by Christal (1971) has been expanded and rewritten in ASCII. The present system, including CURVES, is described in detail in users' manuals stored on computer files at AFERI, Brooks AFB, TX.

The non-SME groups rated surprisingly accurately. This is explained by the 'understandability' of most of the 60 tasks in relation to the benchmark tasks. Overall the mail survey procedure showed merit as SMEs were able to produce accurate, although slightly inflated, ratings.

PHASE II

Method

Eight AFSs were selected for Phase II. The same selection criteria were used as in Phase I. Two AFSs were selected from each of the four aptitude areas of general, administrative, electronics, and mechanical. Of these two, in each aptitude area, one had a high aptitude requirement and one had a low aptitude requirement as detailed in AFR 39-1, Airman Classification. Those selected were 251X0, Weather; 272X0, Air Traffic Control; 304X0, Wideband Communications Equipment; 427X0, Machinist; 542X1, Electric Power Line; 602X0, Vehicle Operator/Dispatcher; 702X0, Administration; and 732X0, Personnel.

The mailable survey instrument for Phase II consisted of four items: (a) a motivational cover letter and instructions as in Phase I; (b) a more detailed background information form than used in Phase I; (c) either an electronics, mechanical, or general/administrative benchmark rating scale booklet, dependent on the aptitude area of the AFS in question; and (d) the list of 60 tasks to be rated on a rating form, these being identical to those originally used by the contractor. Only 40 tasks were used in AFSC 542X1 as was used by the contractor. Random selection of 100 SMEs from the 7-skill levels (and 9-skill levels, where available) of each of the eight AFSs was made. In addition the same non-SME respondents as in Phase I were again selected to complete the survey for AFSC 251X0, Weather.

Analyses of the data collected in Phase II had three major goals. These were to (a) determine the internal consistency of the 25-point benchmark task difficulty ratings for the various groups of raters; (b) determine the validity of the rater groups using the same validity measuring procedures as in Phase I; and (c) compare measures of aptitude specific OLD as in Phase I and non-aptitude specific OLD (Ramadge, 1987) generated from SME benchmark data with those generated from contractor data.

Results

The percentage of useful responses ranged from 22 to 48 across AFSs studied. Reliability of ratings was assessed using the Apollon index (the GRPREL (originally REXALL) (Christal and Weissmuller, 1978)). This is an index of interrater agreement. Reliability indices were derived from the rated group (Table 1). Interrater reliabilities were all above .60, except the two AFSs with lowest numbers of useful responses. The useful responses rate and lowest interrater reliabilities in the AFSs 702X0 and 732X0 are due to three things: (a) there had been significant movement of personnel out of these career fields recently, increasing the number of inexperienced raters; (b) within these career fields there are numerous positions and many of the raters surveyed were not familiar with the tasks in the benchmark; (c) SMEs were given the option not to rate tasks that they were not familiar with. When combined with the first two reasons, this resulted in a low average number of tasks rated per rater.

Correlations among SME, contractor, and supervisor ratings (Table 3) were calculated in the same way as in Phase 1. Correlations between SME ratings and contractor ratings ranged between .79 and .94, except for 542X1 (-.08) and 732X0 (.65). The SME ratings all correlated extremely well (between .83 and .94) with the supervisor ratings. In every case, SME ratings correlated better than contractor ratings with the supervisor ratings. With this in mind, the validity of the contractor ratings for the 542X1, and to a lesser extent, the 732X0 studies may be open to question.

Table 2. Interrater Reliabilities

AFSC	No. of Useful Responses	Interrater Reliability
251X0	40	.98
272X0	41	.97
304X0	34	.98
427X0	49	.98
542X1	35	.93
603X0	33	.96
702X0	27	.92
732X0	22	.88

Table 3. Correlations Between Rating Groups

AFSC	SME vs. Contr	SME vs. Supvs	Contr vs. Supvs
251X0	.83	.91	.82
272X0	.79	.87	.75
304X0	.94	.91	.88
427X0	.90	.85	.79
542X1	-.08	.94	-.09
603X0	.81	.89	.86
702X0	.84	.89	.80
732X0	.65	.83	.65

Comparing SME and contractor average task difficulty ratings revealed higher ratings by the SMEs in every case except 427X0. The average SME rating for 427X0 was 12.55, slightly below the contractor average of 13.19. When examined by aptitude area, all General/Administrative AFS SME average ratings were above the contractor average ratings by between 1.05 (732X0) and 1.94 (272X0). For all Electronic or Mechanical AFSs, the SME ratings varied from .64 below (427X0) to 1.10 above (542X1) the contractor ratings. To investigate the effect of this rating inflation, aptitude specific OLDs were determined using the same procedure as in Phase 1. Common scale OLDs (Ramadge, 1987) were calculated from the aptitude specific OLDs (Table 4). Aptitude specific OLDs based on SME data were inflated by between 4 and 16 for all AFSs, except 427X0 which saw a reduction of 8. Despite this variation, the order on the common scale of the OLDs across the 8 AFSs was maintained, except for 427X0.

Table 4. Average Task Difficulty Ratings and Occupational Learning Difficulties

AFSC	Contractor Avg	SME Avg	Aptitude Specific OLD		Common Scale OLD	
	(60 tasks)	(60 tasks)	Contractor	SME	Contractor	SME
251X0	11.82	13.60	107	123	98	111
272X0	11.86	13.80	107	122	98	111
304X0	13.38	13.85	129	134	121	126
427X0	13.19	12.55	117	109	123	118
542X1	9.92	11.02	99	108	99	105
603X0	7.57	7.97	79	83	88	92
702X0	8.27	9.38	72	80	88	96
732X0	10.26	11.31	87	99	91	97

The non-SME groups who rated the 251X0 survey correlated reasonably well with the contractor data, between .65 and .79. (Data is not shown herein due to the limit on the length of the paper.) The non-SMEs did have greater difficulty in finding the correct level on the scale as their average ratings were between .65 and 2.49 above the contractor ratings. The SMEs did a better job.

CONCLUSION

The mail survey procedure tested in this research proved successful as SMEs were able to reproduce accurate, although in general slightly inflated, estimates of OLD. It should now be possible to produce a translation formula to enable the implementation of this method for collecting accurate benchmark task difficulty ratings. A more complete treatment of these data may be found in a Technical Paper currently in preparation titled "A Practical Procedure for Determining Occupational Learning Difficulty" by this author.

REFERENCES

- Air Force Regulation 39-1. (1982). Airman classification regulation. Washington, DC: Department of the Air Force.
- Burtch, L. D., Lipscomb, M. S., & Wissman, D. J. (1982). Aptitude requirements based on task difficulty: Methodology for evaluation (AFHRL-TR-81-34, AD-110 568). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Christal, R. E. (1974, January). The United States Air Force occupational research project (AFHRL-TR-73-75, AD-774 574). Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory.
- Christal, R. E., & Weissmuller, J. J. (1976). New CODAP programs for analyzing task factor information (AFHRL-TR-76-3, AD-A026 121). Lackland AFB, TX: Occupational and Manpower Research Division, Air Force Human Resources Laboratory.
- Garcia, S. K., Ruck, H. W., & Weeks, J. (1985). Benchmark learning difficulty technology: Feasibility of operational implementation (AFHRL-TP-85-33, AD-A161 797). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Mead, D. F., & Christal, R. E. (1970). Development of a constant standard weight equation for evaluating job difficulty (AFHRL-TR-70-44, AD-720 255). Lackland AFB, TX: Personnel Division, Air Force Human Resources Laboratory.
- Ramadge, J. D. (1987). Task learning difficulty: Interrelationships among aptitude-specific benchmarked rating scales (AFHRL-TP-86-56). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Weeks, J. L. (1984). Occupational learning difficulty: A standard for determining the order of aptitude requirement minima (AFHRL-TP-84-76, AD-147 410). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

JOB INCUMBENT PERFORMANCE REPORTED BY
ABSOLUTE FREQUENCY - A FURTHER EXAMINATION

LAWRENCE A. GOLDMAN, Ph.D.

U.S. ARMY SOLDIER SUPPORT CENTER - NCR

Background. Since the adoption of the Comprehensive Occupational Data Analysis Programs (CODAP) by the Army Occupational Survey Program (AOSP) in the early 1970's, job incumbent information collected from enlisted soldiers has been based on a seven (7) point Relative Time Spent (RTS) scale. Before using this scale, job incumbents are first asked to review all tasks in the occupational questionnaire. Then they are asked to rate each of the tasks which they perform on their current job using the RTS scale, based on the amount of time devoted to that task relative to all other tasks which they perform. From a theoretical standpoint, this approach is fine. However, from a practical standpoint it ignores the obvious resistance of the individual soldier in going through the entire task list twice (AOSP questionnaires usually contain between 300 and 400 tasks).

While each of the values for the RTS scale is defined on each page of the "tasks" section of the AOSP questionnaire, ranging from (1) "very much below average time spent" to (7) "very much above average time spent," the unmodified RTS scale values cannot be used by themselves. To illustrate, a common job description could be obtained from two individuals performing the same tasks, although one rated these tasks consistently lower than "average time spent" while the other rated these tasks consistently higher than "average time spent." The data collected through use of the RTS scale have been found to be most useful for: 1) assisting in job "typing" (pertaining to the identification or classification of clusters of individuals formed on the basis of similarity of work performance); and 2) examining the work performed by sub-groups of individuals based on the relative amount of time spent on each duty (a duty representing a cluster of closely related tasks). In contrast, the information obtained from average percentage time spent values reported by job incumbents at the task level cannot be used readily for critical task selection. Consequently, several U.S. Army service schools requested that a frequency scale in addition to, or as replacement to, the RTS scale be used for surveys of enlisted Military Occupational Specialties (MOS) for which they are proponents. An Absolute Frequency (AF) version of the frequency scale was thought to be desirable for this objective, primarily because it would be easier for job incumbents to rate each task performed since each rating would be independent, regardless of the number of tasks performed. Nonetheless, if a frequency scale were to be used for occupational surveys, it should have, at a minimum, the same capabilities as the RTS scale with respect to job typing and identification of work at the duty and task level. Regular use of an AF scale for Army-wide surveys would be feasible if tests could indicate a close comparability between the AF scale and the RTS scale in these respects. This effort comprises a follow-up to a previous study of the RTS and AF scale I reported upon at the 1986 Military Testing Association entitled "Examination of the Feasibility of an Absolute Frequency Scale." In this other study, MOS 12B (Combat Engineer), MOS 34K (Combat Signaler), and MOS 94B (Food Service Specialist) were examined.

Methodology. The results reported in this study were based on Army-wide sample surveys of job incumbents in: 1) MOS 13B, Cannon Crewman, with authorized skill levels (DA) 1, 2, 3, and 4 corresponding to paygrades E-3/4, E-5, E-6, and E-7; 2) MOS 34M, Multichannel Communications Systems Operator,

authorized SL1, SL2, and SL3; and 3) MOS 31Q, Tactical Satellite/Microwave Systems Operator, also authorized SL1, SL2, and SL3. In October 1987, 1,184 MOS 13B questionnaires using the RTS scale and 1,168 questionnaires using the AF scale were randomly distributed to soldiers at the same duty locations Army-wide. In November 1987, 670 MOS 31Q questionnaires with the RTS scale and 668 MOS 31Q questionnaires using the AF scale were administered Army-wide. Also in November, 1,680 MOS 31M questionnaires containing the RTS scale and 1,682 MOS 31Q questionnaires using the AF scale were administered Army-wide. Analysis of the data was based on the following sample sizes:

	<u>MOS 13B</u>		<u>MOS 31Q</u>		<u>MOS 31M</u>	
	<u>RTS</u>	<u>AF</u>	<u>RTS</u>	<u>AF</u>	<u>RTS</u>	<u>AF</u>
SL1	379	373	175	154	691	666
SL2	87	85	48	52	167	148
SL3	55	58	13	12	46	47
SL4	<u>11</u>	<u>9</u>	<u>n/a</u>	<u>n/a</u>	<u>n/a</u>	<u>n/a</u>
Totals	532	525	236	218	904	661

Each survey contained a list of tasks specific to the MOS grouped into duty areas. The duty/task section of the MOS 13B survey comprised 485 tasks arranged into 20 duties; this section for MOS 31Q and MOS 31M, included within a common questionnaire booklet, contained 365 tasks in 16 duty areas. Each of these three MOS was equipment-specific. In contrast, none of the three MOS surveyed previously (MOS 12B, 31K, and 94B) was based primarily on operation/maintenance of specific items of equipment.

The seven (7) values for the AF scale used in these surveys were: 1) "less often than once a month;" 2) "once a month;" 3) "2 or 3 times a month;" 4) "once a week;" 5) "2 or 3 times a week;" 6) "once a day;" and 7) "more often than once a day." In this format, the AF scale was simply a nominal scale which greatly limited its usefulness since it was not feasible to compute average frequency values. Therefore, it was decided to use an estimated measure of yearly frequency which would have the overall effect of changing this nominal scale to an interval type of scale. This measure pertained to the concept of "frequency per year," which assumed that a typical soldier worked 5 days a week (regardless of the number of hours worked per day) for 11 months of the year (assuming 30 days of annual leave). The value label "less often than once a month" was interpreted as four (4) times a year (that is, once every month) while "more often than once a day," was designated as twice a day. Each of the seven AF values was then converted as follows, corresponding to estimated frequency of performance per year: 1) "4;" 2) "11;" 3) "28;" 4) "52;" 5) "120;" 6) "240;" and 7) "480." The traditional Army RTS scale was used, ranging from 1) "very much below average," to 7) "very much above average." The following analyses were conducted for MOS 13B, 31Q, and 31M using COMAP:

1. Identification of information contained at the duty/task level. To test for comparability at the duty level, rank orders based on average percent of members performing within each duty area, for each skill level, were determined for RTS and AF scale data. Rank order equivalency was then determined using the Spearman Rank-Order Correlation Coefficient (Spearman's

rho). This approach, done for the first time, examines duty level performance in terms of a measure (task performance) common to the RTS and AF scale questionnaires. Also, for the first time, task level Pearson product-moment correlations between average percent time spent values (RTS scale data) and average estimated yearly frequency values (AF scale data) were determined for the skill levels of each MOS. A further test, as done in the previous study, was based on the belief that the average number of tasks performed by all job incumbents within a MOS should be closely comparable regardless of the type of scale used in the survey.

2. Identification of major job types. The process used within CODAP is based on "clustering" or "hierarchical" principles by which individuals are grouped reflecting similarity of work performed based on task performance ratings for: (a) commonality of time spent values (RTS scale); or (b) estimates of average frequency values (modified AF scale). It was hypothesized that if these two rating scales could be used interchangeably for the purpose of job typing, then the AF scale should "capture" the same job types as those found using the RTS scale. Job types comprising at least five (5) percent of the total sample, regardless of scale, were noted. As an additional test for comparability between these two scales, differences in terms of percent of members performing were cited. These analyses were also conducted in the previous study.

3. Analysis of the distribution of task value responses. For each of these scales, the average number of job incumbents rating each task value was determined. It was hypothesized that for the RTS scale, the distribution should be roughly in the form of a bell-shaped curve, with the highest average value being in the middle (4 - "average time spent"), with approximately the same number of raters rating tasks below average as above average. With respect to the AF scale data, however, there was no a-priori hypothesis concerning the distribution of responses. This analysis was also done in the previous study.

Findings.

A. Comparisons of Job Information at the Duty/Task Level. Regardless of the method used to determine comparability at the duty or task level between RTS and AF scale respondents for MOS 13B, 31Q, and 31M, the analysis indicated extremely close comparability. With respect to examination of rank order correlations based on the average percentage of RTS scale and AF scale job incumbents by duty area for each skill level, the correlations were remarkably high for all these MOS, the Spearman's rho being .92 or greater for each SL. With respect to the Pearson product-moment correlations between average percent time spent (RTS scale) and average estimated yearly frequency (AF scale) task values by skill level, these correlations were also all highly significant. By skill level, these correlations were .91, .93, and .92 for SL1 13B, 31Q, and 31M job incumbents, respectively. For SL2 soldiers in these three MOS, the correlations were .81, .90, and .86, respectively; for SL3 soldiers, they were .77, .61, and .83. For MOS 13B soldiers, the SL4 correlation was .71. In terms of the overall average number of tasks performed, again there was close comparability for each of these three MOS. For MOS 13B, the average of 131.4 tasks performed by RTS scale respondents was essentially identical to that of their AF scale "peers" - 131.5. For MOS 31Q and 31M, the average number of tasks performed by RTS scale respondents (94.1 and 93.3) was just slightly higher than that indicated for AF scale users - 87.8 and 89.0 for MOS 31Q and 31M, respectively.

B. Job Structure Analysis. The job typing for MOS 13B with respect to RTS scale respondents yielded eight (8) distinct job types, comprising 500 incumbents (94.0 percent of the total sample). While these same job types were also found for MOS 13B AF scale respondents, there were three (3) additional job types that were identified only for AF scale respondents. The 11 job types identified for the AF scale comprised 490 soldiers, representing 93.3 percent of the total AF sample. There were five job types common to both RTS and AF scale users, significant in that each comprised at least 5 percent of the total number of RTS scale or AF scale incumbents whose job was identified: 1) Towed Howitzer Specialists; 2) Self-Propelled Howitzer Specialists; 3) Self-Propelled Howitzer/Tracked Cargo Carrier Operators; 4) Self-Propelled Howitzer/Nuclear Cannon Assembly Operators; and 5) Howitzer Section Chiefs. Another job type, Nuclear Cannon Assembly Specialists, was isolated only for respondents using the AF scale. All these job types were nonsupervisory in terms of the primary type of work performed with the exception of Howitzer Section Chiefs. A test of comparability between "common" job types was done for the two largest nonsupervisory job types (Towed Howitzer Specialists and Self-Propelled Howitzer Specialists) and the one supervisory job type. It was found that differences were minor for the nonsupervisory job types in terms of percentage of members performing. For Towed Howitzer Specialists, there were only 26 tasks (5.4 percent of the 485 tasks in the 13B questionnaire) for which the percentage of soldiers performing differed by at least 20 percent. For Self-Propelled Howitzer Specialists, there were just 40 tasks (8.2 percent of all tasks) indicating significant differences. However, with respect to Howitzer Section Chiefs, there were 108 tasks (22.3 percent of all tasks) reflecting significant differences, primarily involving hands-on, Nonsupervisory responsibilities. Of these 108 tasks, all but five related to significantly fewer percentages of Howitzer Section Chiefs responding to the AF scale.

The job typing for MOS 31Q relating to these two scales yielded six common job types, all comprising at least 5 percent of the total number of RTS or AF scale job incumbents identified. These were: 1) Tactical Satellite/Microwave System Operators; 2) Tactical Microwave Systems Operators (Generalists); 3) Tactical Microwave Systems Operators (Radio Terminal Set Specialists); 4) Tactical Microwave Systems Operators (Radio Repeater Set Specialists); 5) Tactical Microwave Team Chiefs; and 6) Tactical Satellite Systems Operations Team Chiefs. With respect to just AF scale respondents in MOS 31Q, an additional job type was noted -- Signal Security Specialists. Of the 236 individuals in the MOS 31Q RTS scale sample, 208 (88.1 percent) were job typed. In the corresponding MOS 31Q AF scale sample, 205 (94.0 percent) of the incumbents were classified. Based on the criterion of task performance, there were very few, if any, significant differences between nonsupervisory RTS and AF scale respondents in MOS 31Q. For example, there were no differences exceeding 20 percent based on the percent of Tactical Satellite/Microwave System Operators performing. For the supervisory job types (Tactical Microwave Team Chiefs and Tactical Satellite Systems Operations Team Chiefs), there were substantially greater differences. For example, there were 57 tasks (15.6 percent of all tasks) reflecting significant task performance differences for Tactical Microwave Team Chiefs. Of these 57 tasks, just slightly less than half pertained to significantly fewer percentages of these team chiefs using the AF scale.

With respect to MOS 31M, there were 12 job types common to both RTS and AF scale respondents. Of these 12, only five represented 5 percent or more of

either the total RTS or the AF scale sample. These were: 1) Multichannel Communications Operator (Generalists); 2) Multichannel Communications Operators (AN/TRC-145(V) Radio Terminal Set Specialists); 3) Multichannel Communications Operations Team Chiefs; 4) Multichannel Communications Team Chiefs; and 5) Multichannel Communications Section Chiefs. In addition, there were two job types isolated only for AF scale respondents - COMSEC Account specialists and Signal Security Specialists. On the other hand, there was a job type - Multichannel Communications Operators (AN/TCC-73 Telephone Terminal Set Specialists) that was isolated only for 31M RTS scale respondents. Of the 904 individuals in the RTS scale sample, 814 (90.0 percent) were job typed. Pertaining to the 31M AF scale sample, 620 (93.8 percent) were classified. Similar to the other MOS in this study, there were virtually no differences between the nonsupervisory 31M job types in terms of task performance. In particular, there were no significant task differences for the "Generalists" and only seven tasks (1.9 percent of all tasks) for which differences exceeded 20 percent for Multichannel Communications Operators (AN/TRC-145(V) Radio Terminal Set Specialists). On the other hand, there were 65 tasks (17.8 percent of all 365 tasks) for which significant differences were noted for supervisory Multichannel Communications Section Chiefs. Of these 65 tasks, all but three pertained to significantly greater percentages of these section chiefs responding to the AF scale.

In the previous study of MOS 12B, 31K, and 94B, identification of the job types in the AF scale samples was possible, for the most part, on the basis of very few tasks. However, for these three (equipment-oriented) MOS, job typing was based on approximately the same number of tasks for both the RTS scale and the AF scale samples. Another difference between the current and previous study was that in the previous one, there were no clear-cut differences between the common nonsupervisory and supervisory job types in terms of percent of soldiers performing tasks.

C. Average Number of Job Incumbents Responding to Each Task Value. The average number of MOS 13B, 31Q, and 31M job incumbents responding to each value of the RTS and the AF scales was obtained to examine response distribution. With respect to RTS scale data for MOS 31Q and 31M, the average number of incumbents responding to values (1) and (4) - "very much below average" and "average time spent", respectively, - was substantially higher than the other values. For MOS 13B incumbents using the RTS scale, the only clear-cut finding was that the average number of individuals using the middle value "average time spent" was appreciably higher than the other six values. With respect to AF scale data for MOS 13B, the average number of job incumbents responding to each scale value decreased monotonically from the first scale value through the seventh scale value. The same finding was noted for MOS 31Q and 31M AF scale respondents except that the second highest average was reported by soldiers performing "once a week" (the fourth scale value). Because of this exception, MOS 31M and MOS 31Q RTS and AF scale respondents were similar to each other in that the highest average number of individuals rated either the first or fourth scale values. It was believed that while very few tasks are performed on a recurring basis by MOS 13B Cannon Crewman, MOS 31Q and 31M soldiers are likely to operate and/or perform operational maintenance regularly (e.g., once a week) on the radio/telephone equipment for which they are primarily responsible, or infrequently on other types of equipment. These findings were also found in the previous study of MOS 13B, 31K, and 94B, with the exception, pertaining to AF scale users, that the average number of job incumbents invariably decreased from the first (most infrequent) scale value to the seventh (most frequent) scale value.

Conclusions and Implications for Future Studies. It was evident that there was a substantially high degree of comparability between the RTS scale and the AF scales based on the analyses of equipment-specific MOS 13B, 31Q, and 31M, just as there was in the previous study of non-equipment specific MOS 12B, 31K, and 94B. This was particularly true in terms of the correlational studies at the duty and task level, done for MOS 13B, 31Q, and 31M, that were not previously accomplished. In terms of the average number of tasks performed overall, it was observed that the differences were insignificant for MOS 13B, 31Q, and 31M whereas, in the previous study, significantly fewer tasks were performed by MOS 12B, 31K, and 94B job incumbents using the AF scale.

Insofar as job structure analysis was concerned, it was noted in the current study, just as it was for the previous study, that the job types which were identified through the use of the "traditional" RTS scale also emerged, essentially without exception, through use of the "non-traditional" AF scale. Also, use of the AF scale appeared to identify job types that were not identified for RTS scale job incumbents. One new finding, based on the analysis of MOS 12B, 31Q, and 31M, was that there are apparent differences in terms of probability of performance between supervisory and nonsupervisory job types. With respect to the nonsupervisory job types, there were few, if any, significant differences. However, with respect to supervisory job types, task performance differences seemed prominent. This latter finding could be due, at least partially, to the substantial degree of diverse nonsupervisory work performed by supervisors in these three MOS.

Another key finding was that the correlations based on average percent time spent (RTS scale) and average estimated yearly frequency (AF scale) were highest for SL1, decreasing monotonically from SL1 through SL3 for MOS 31Q and 31M, through SL4 for MOS 13B. This could be explained, at least in part, by the fact that supervisory tasks performed at the upper SL tended to be "soft skill", less likely to be time ratable as well as ratable in terms of frequency of performance.

As a further validation study, three MOS are currently being surveyed using split samples of both the RTS and the AF scale. One of these MOS is equipment-specific while the other two are not. The equipment-specific MOS 31C, Single Channel Radio Operator, is quite close to MOS 31Q and 31M. The others are MOS 71L, Administrative Specialist, and MOS 91A, Medical Specialist. The results for these three additional MOS will be available in 1989. After analyses of these three new MOS, it will be of interest to ascertain the extent to which the results for each MOS relate to being equipment-specific. It seems clear at this point, however, that there is merit in considering widespread use of occupational surveys using the AF scale to provide more useful information for U.S. Army training school course/combat developers and evaluators.

CODAP: Organizing Operational Applications and R&D

Outside of the U.S. Military

by

Michael R. Staley
and
Johnny J. Weissmuller

of

Sensible Systems, Inc.

BACKGROUND

Within the United States Air Force, as well as many other military and civilian agencies, there is a fundamental technology which supports both the operational and research occupational analysis programs. The core of this technology is called CODAP, an acronym for the Comprehensive Occupational Data Analysis Programs. The CODAP "system" is a set of analysis tools and procedures which use, as raw material, information provided by the members of the occupational field being studied. This system may be used to revise classification structures, assess job related skills, verify the relevance of training courses and a host of other applications in which an accurate knowledge of job content at the task level is desirable (Morsh and Christal, 1966; Christal, 1974; Levine, Ash, Hall and Sistrunk, 1983; Christal and Weissmuller, 1988).

During their years associated with the U.S. Air Force, the principals in Sensible Systems, Inc. tried to "give away" the CODAP system to appropriate non-profit agencies as part of a federal transfer of technology program. Although the CODAP system took root in other military organizations and a few universities, it withered and died out in most state and local governmental agencies (Goodison, 1980; Frank, 1981). After leaving the Air Force, the principals in Sensible Systems, Inc. formed this corporation to create a new version of CODAP. This new version would be built from the ground up and would provide the best features of the mainframe CODAP systems and still be of use to state and local governments.

In November 1986, Sensible Systems, Inc. announced the availability of a commercial service bureau to provide CODAP analysis products from client provided data. The scope and capability of this "microcomputer" based package which incorporates the

most unique data processing capabilities of the mainframe CODAP systems was reported at the May 1987 Occupational Analysts Conference hosted by USAFOMC (Staley, Weissmuller, Lewis and Johnson, 1987). This new package is called "atCODAP", which means "anchored to CODAP, but not limited by it."

INTRODUCTION

The purpose of this paper is to report on two years of experience dealing with clients who were not fully "CODAP literate." In the first section of the paper we will deal with many of the misconceptions which creep into a CODAP job analysis when directed by someone with a non-CODAP background. In the second section of the paper we will talk about the co-operative research Sensible Systems, Inc has underwritten with the co-operation of our clients. Finally, in the conclusion we will point out the direction in which we are headed.

WHAT CAN GO WRONG IN A CODAP PROJECT?

The Task List

Fortunately, this paper is limited to just six pages, so we will just touch on the highlights. One major difference between military and non-military users of CODAP is the willingness to accept "THE TASK LIST" as the cornerstone of all good job descriptions. The military view, in fact, is to talk about "TI/CODAP" which means the "Task Inventory (processed by) CODAP." This viewpoint has its merits and its drawbacks. The primary merit is that it establishes CODAP as a Behaviorally Anchored Job Analysis System or CODAP/BAJAS. Note that CODAP was originally developed in San Antonio and with that Spanish influence "Baja" can be interpreted to mean "beneath" or "underlying." As psychologists, we should remain aware of the dual level of behavior being discussed here - the surface level of self-reported behavior as well as the underlying verbal behavior required to recognize a printed statement as relevant to the respondent's job. One can always "spot" an inventory developed by a person from a "testing" background - they incorporate "clever" lie scale items which will have to be removed prior to processing. They do not understand that the programming required to remove lie scales probably introduces more errors than the few careless people they "trapped."

Where is all this leading? Well, if one fails to grasp the point, in several wrong directions. First, the "task statement" should not be considered the primary backbone of the CODAP system. Because other agencies want other lists to play important roles in their decisionmaking processes, we have seen Knowledge, Skills, and Ability (KSA) lists as well as equipment lists literally stuck on the end of the task list - using the same 5-point Likert spent rating scale. In mainframe CODAP, a KSA list (with an appropriate rating scale)

can be submitted as a second study (KSA/CODAP instead of TI/CODAP) with the resulting confusion making it hard to "coordinate" these parallel studies.

Although micro-based "atCODAP" supports the standard task list approach, it does NOT take a TASK LIST as a primary entity in its analysis outlook. Up to 99 different lists may be defined and used within a single analysis. As many as 999 data factors may be collected for each list in a study. Lists may include such things as Knowledges, Skills, and Abilities (KSAs); Equipment Used; Equipment Maintained; Job Satisfaction Items; Courses Taken; etc. This multiple list capability allows you to perform a clustering based upon data from one list which produces a set of cluster stages that form groups of people. Job descriptions for these groups may then be reported using any list with its associated factors.

One fundamental assumption of every version of the CODAP technology is that the most important data are collected directly from the job incumbent. In order to ensure the validity of responses it is necessary to use the language of the job incumbents in the survey instrument. Language used by psychologists in the analysis of the information collected should be avoided in the survey instrument unless it is directly used in the work environment (e.g. Mental Health Clinic Technicians). We have seen examples where this rule has been violated such as a KSA item: "Mentally organize a disorganized field into a single picture (Closure)."

The Rating Scale

In mainframe CODAP one has the Task List and Background Information. Many "lists" may be buried in the background information section such as equipment lists, courses taken, etc. Because of the limited reporting capability of mainframe CODAP, these items are typically rated "Use/Don't Use", or some equally binary format that can be coded "0" (zero) for no and "1" (one) for yes. When the 99 list possibility of atCODAP is utilized, users sometimes overlook the possibility of clustering based upon a given list and assign scale points which are nearly impossible to use statistically.

The example which comes most clearly to mind is that of a Knowledge, Skills, and Abilities list. In the example we are referring to, this list was presented twice. The first presentation asked for ratings of the KSA's relevance to successful job performance (0=not related, 1=helpful, 2=important, and 3=essential). The second presentation asked for ratings of the value of having the KSA at entry into the job (0=not required, 1=helpful, 2=important, 3=essential). Because these scale points represent such diverse values, it would be inappropriate to compute job descriptions, let alone convert it into a scale for clustering. To serve as a good basis for a job description or a clustering, a rating scale should reflect evenly spaced alternatives as points on the scale. Although mainframe CODAP promotes the use of a nine point scale most agencies find a seven point scale adequate for their needs.

The Inventory Length

Traditional CODAP dictates that we collect job information from the job incumbent because he or she knows the job best. This implies a certain respect for and trust in the job incumbent. Because we believe they are proud of their work, we want to keep them away from their work for as a short period as possible. Traditionally, we try to design an inventory such that a job incumbent can complete it within two hours. Unfortunately, with some of the flexibilities we've designed into atCODAP, it becomes easier to make surveys which cannot be completed within four hours. Inventory designers must keep in mind the use they expect to make of each data element as it goes into the survey. "Nice to have", but not essential information could be the straw that breaks the camel's back.

It is an important consideration to keep in mind when exploiting the atCODAP multiple list capability is that humans will need to respond to the survey instrument that is developed.

RESEARCH UNDERWRITTEN BY SENSIBLE SYSTEMS

City of Fort Worth

Although the analysis of this project is not yet complete, Sensible Systems, Inc. and the City of Fort Worth undertook a project to evaluate the effects of collecting task factors in a side-by-side format versus three separate, single-factor booklets. It is our expectation that not only will the interrater reliabilities be higher for the single factor booklets, but also the correlation between the "independent" factors will be lower. Because the multi-factor booklets were administered in a proctored classroom setting we expect this to represent the best possible multi-factor scenario.

Ontario Hydro Electric

Many years ago (Haltrecht, 1980) Ontario Hydro (OH) was one of CODAP's early proponents. In the ensuing years, however, those with special CODAP skills and knowledge moved on to other positions. So, when Ontario Hydro found itself in need for CODAP analyses, it found that the computer people required to use the software had all gone and no technical capability was left. To produce some product within the given schedule, an inventory was developed and administered to job incumbents. Ontario Hydro had problems when it came time to cluster the results. The standard statistical packages they had access to could not handle the specified volume. Sensible Systems agreed to cluster their data at no cost.

When the data arrived, we were surprised at the inventory booklet. This was not a standard relative time spent booklet, but a survey rated on approximate frequency (1 =

YEAR -- on a five point scale. Unable to reach the OH contact point, we performed the clustering twice, once on the five point scale and one on a substitution scale which spanned from 1 to 240.

We forwarded the products with a explanation of each set and asked that the subject-matter-experts be shown both sets. As per our expectations, the experts preferred the substitution scale results which then formed the basis of their other analyses.

City of Minneapolis

The City of Minneapolis, and their outstanding consultant, Dr. Gail Drauden, have been very cooperative in joint undertakings. We have, with Dr. Drauden's guidance, developed methodologies for clustering composite job descriptions in a manner similar to the way in which we cluster individual job descriptions. Dr. Drauden was interested in this technique in order to expedite the validation of promotion/selection tests.

FUTURE DIRECTIONS

Sensible Systems intends to maintain its status as a commercial service bureau capable of providing CODAP support. We also intend to publish a CODAP analysts manual to help forestall the kinds of problems we have encountered in the past two years. Sensible Systems, Inc. was not founded as a non-profit corporation, it just seems to have turned out that way. To try to rectify this situation, we are currently exploring two possible avenues for continuation of any non-profit activities. The first possibility is a associated with a university. The second non-profit organization is a new corporation which will become active in January 1989. The purpose of this second agency is to administer training programs under the Joint Training Partnership Act (JTPA). In either case, we are looking for a home for the public domain job inventories we are collecting as well as a home for a newsletter to keep occupational analysts apprised of what's going on in their field.

BIBLIOGRAPHY

- Christal, R. E. (1974). The United States Air Force occupational research project (AFHRL-TR-73-75, AD 774 575). Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory.
- Christal, R. E. & Weissmuller, J. J. (1988). 9.3 The Job-Task Inventory. in The Handbook of Job Analysis. Sidney Gael (Ed). John Wiley & Sons. New York. (Volume 2)
- Frank, M. S. (1981). The Philadelphia Experience: A Critique and analysis of the applicability of the Comprehensive Occupational Data Analysis Programs (CODAP) for civilian personnel management. Draft Report for The National League of Cities Service Program on Employer-Employee Relations
- Goodison, M. (1980). A summary of 10 cities on-site interviews and a critique of CODAP (Comprehensive Occupational Data Analysis Programs) Report: Center for Occupational and Professional Assessment. Princeton NJ: Educational Testing Service
- Haltrecht, E. (1980). CODAP: Introduction and uses in a large public utility. Proceedings 22nd Annual Conference of the Military Testing Association HAI-0-15. Toronto, Ontario Canada: Military Testing Association.
- Levine, E. L., Ash, R. A., Hall, H., & Sistrunk, E. (1983). Evaluation of Job Analysis Methods by Experienced Job Analysts. Academy of Management Journal, 26(2), 339-348.
- Morsh, J. E. & Archer, W. B. (1967). Procedural Guide for Conducting Occupational Analysis in the United States Air Force (PRL-TR-67-11, AD-664 037). Lackland AFB, TX: Personnel Research Laboratory.
- Morsh, J. E. & Christal, R. E. (1966) Impact of the computer on job analysis in the United States Air Force (PRL-TR-66-19, AD-656 304). Lackland AFB, TX: Personnel Research Laboratory, Aerospace Medical Division.
- Phalen, W. J., Weissmuller, J. J., & Staley, M. R. (1985). Advanced CODAP: New Analysis Capabilities Proceedings Fifth International Occupational Analysts Workshop. San Antonio, TX: USAF Occupational Measurement Center. (May)
- Staley, M. R., Weissmuller, J. J., Lewis, T. D., & Johnson, C. A. (1987). atCODAP: A Definition. Proceedings of the Sixth International Occupational Analysts Workshop 149-156. San Antonio, TX: USAF Occupational Measurement Center. (May)
- Staley, M. R., Weissmuller, J. J., & Phalen, W. J. (1985). ASCH CODAP: The Impact of a System Designed for Emerging Applications. Proceedings of the Fifth International Occupational Analysts Workshop Randolph AFB TX: USAF Occupational Measurement Center. (May)
- Weissmuller, J. J. (1979). CODAP: Applications for state and local governments. (Exhibit) Harvard-CIT Federal Technology Transfer Forum and Exposition.
- Weissmuller, J. J., Moore, B. E., & Thew, M. C. (1980). CODAP: Applications and their implications for higher level design. Proceedings 22nd Annual Conference of the Military Testing Association WEI-1-5. Toronto, Ontario Canada: Military Testing Association
- Weissmuller, J. J., Staley, M. R., Lewis, T. D., & Johnson, C. A. (1987). atCODAP: Covered Prices. Municipal Prices. Proceedings of the Sixth International Occupational Analysts Workshop 1-166. San Antonio, TX: USAF Occupational Measurement Center. (May)

A COMPARISON OF METHODOLOGIES FOR GROUPING LARGE NUMBERS OF OCCUPATIONS .

by John E. Stein, Donald E. McCauley, and Brian S. O'Leary
U.S. Office of Personnel Management
Washington, D.C.

INTRODUCTION

The purpose of this study was to explore methods of grouping professional and administrative occupations in terms of the similarity of the work performed.

Developing examinations for individual occupations is always expensive and, depending on the number of hires into each occupation, may not be cost-effective. Grouping jobs on the basis of work behaviors provides one way of reducing the cost of examination development while not sacrificing test validity. This is a common approach. Barnes and O'Neill (1978) grouped jobs for examination development in the Canadian Public Service. Gage, Borman, Campbell and Osborn (1983) clustered U.S. Army enlisted jobs into homogeneous groups according to rated job content in order to choose a representative sample of MOS's for test validation purposes.

As Cornelius, Carron and Collins (1979) indicate, there are two distinct decisions that must be addressed in any job grouping or job classification study. The first decision involves selecting the type of job analysis data that will be used to compare jobs. That is, on what attributes of the job will the similarity of jobs be assessed? The second decision involves selecting the type of procedure to use in assessing the similarity of jobs.

Selecting the unit of analysis: There are at least three major units of analysis:

1. Tasks performed - In the most popular job analysis approach the job is broken down into elemental units called tasks. The work of Christal (1974) and his associates typifies this approach.

2. Human behaviors - A second popular approach is that associated with McCornick (1988) and his associates in their work on the Behavioral Analysis Questionnaire (PAQ). In this approach the unit of analysis is not specific tasks but rather the generalized human behaviors required to perform the work (e.g., comparing, synthesizing, planning, etc.).

3. Abilities - A third unit of analysis is the abilities approach. In this approach, most closely associated with the work of Fleishman (1981) and his associates, the unit of analysis is the underlying abilities required to perform the work.

Opinions expressed in this paper are the authors' and do not necessarily represent the official policy of the U.S. Office of Personnel Management.

With method of analysis held constant, at least one published study has indicated that the unit of analysis affects the resultant job groups. Studying foreman jobs, Cornelius and his associates (1979) found that the different units of analysis yielded different results regarding both the number of similar foreman jobs as well as differences regarding which jobs were most similar.

The unit of analysis that we chose for the present study most closely relates to the task approach - but with an important distinction - rather than rate each individual task, the raters were asked to make "holistic" judgments about the similarity of jobs.

Support for the use of this approach comes from the work of Cornelius, Schmidt, and Carron (1981). These researchers compared direct job classification judgments from supervisors and incumbents with the results obtained from a job analysis inventory administered to 1173 job incumbents in 51 plants from 30 companies across the United States. The "holistic" judgments compared quite well with the more elaborate statistical procedure (96% correct classification).

Thus, the "holistic" or whole job approach, which is much less time consuming, yields results similar to more detailed job analyses. Rosse, Borman, Campbell and Osborn (1983) used a whole job approach when clustering U.S. Army enlisted jobs.

Selecting the method of analysis: As indicated earlier, the second decision that must be made in forming job families involves selecting the type of procedure to be used in determining the similarity of jobs. Harvey (1986) has presented an excellent review of the quantitative approaches to job classification, listing the advantages and disadvantages of each. He has classified the procedures into descriptive and inferential methods. Descriptive techniques are those concerned with using an exploratory approach to uncovering groupings of entities. The inferential methods, on the other hand, employ statistical tests of job differences. In the present study we were concerned solely with exploratory methods. We examined three different exploratory methods: factor analysis, cluster analysis, and multidimensional scaling.

METHOD

Data Collection. One hundred thirteen professional and administrative occupations in the civilian, federal work force were studied. First, personnel professionals grouped the occupations into categories according to similarity of work behaviors. These raters were given descriptions of the jobs which were taken from the Federal Government's Dictionary of Occupational Groups and Series of Classes (1964). These descriptions consisted of a brief summary of the job, a list of which summarized the major duties of the job. These job descriptions were printed on 5 x 7 cards and given to the raters for sorting. The General Occupational Series numbers were not included in the cards.

similarities in work behaviors. No limitations were put on the number of categories each rater could generate.

A group of nine raters completed the sort: eight personnel research psychologists and one personnel staffing specialist.

The categories resulting from each of the nine sorts were transformed into a 113 by 113 matrix for each rater wherein a one in a cell indicated that those two jobs were placed in the same category by the rater and a zero in a cell indicated that the two jobs were not placed together. The nine matrices thus derived were added together producing a summary matrix for all nine raters. The values in this matrix ranged from zero (no rater put the two jobs together) to nine (all raters put the two jobs together). This summary matrix was the input for the cluster analysis.

Cluster Analysis (CA). A SAS average linkage cluster analysis was performed on the summary matrix. A seven-cluster solution was chosen due to a sharp drop in the goodness-of-fit index and a small increase in the R-squared value. At this stage of the clustering, all jobs had been merged into the seven clusters and the R-squared value was approximately 70%.

Factor Analysis (FA). The summary matrix was transformed slightly for the FA. The value in each cell was divided by the total number of raters to get the proportion of raters placing both jobs in the same category. Then, an estimate of a correlation coefficient was obtained by taking the square root of the proportion in each cell (Andrews and Ray, 1957). An exploratory Q-type factor analysis was performed on the correlation matrix thus obtained. The FA produced 13 factors which accounted for 99.5% of the variance. In order to make the comparison with the seven-cluster CA solution, the FA was redone limiting the number of factors to seven. These seven factors accounted for 74% of the variance.

Multidimensional Scaling (MDS). The summary matrix was once again transformed for input into the MDS analysis. In this case, the value in each cell in the matrix was subtracted from 10 to produce a dissimilarity or distance matrix, in which larger numbers represented greater distances.

The data were analyzed using the SAS version of ALSCAL. Since the 12,769 cells in the matrix all contained values ranging from one to ten, ties were broken before the matrix was input into the ALSCAL program. This was done by adding or subtracting a value of 1/10,000 from each occurrence of the same value and iterating until there were no more ties. MDS solutions were determined for one through six dimensions using both the metric and nonmetric analysis options. From the stress and R-squared data, the three-dimensional solution appeared to be the most reasonable. In three dimensions, the rank order correlations between the order of the jobs as obtained by the metric and nonmetric solutions were above .9 for each dimension indicating that the two solution methods produced very similar results. In three dimensions, there appeared five tightly grouped clusters and two clusters which were more disparate.

RESULTS

The number of jobs in each cluster as obtained by each analysis method is shown below in Table 1.

Table 1

Number of Jobs in Each Cluster for Each Methodology

	Cluster						
	1	2	3	4	5	6	7
<u>Method</u>							
FA	26	18	19	14	15	12	9
CA	20	18	33	10	13	9	10
MDS	23	19	27	12	10	10	11

As one can see in this table, the number of jobs per cluster was relatively stable across the three solutions with the exception of the third cluster. It should be noted that inclusion of jobs in MDS clusters is more subjective than in the other two methods and that some jobs could reasonably be included in more than one cluster.

Table 2, below, illustrates the agreement across the three analysis methods by providing the number of jobs in common between each pair of solutions. The percentage of jobs in which the three solutions agree was 83%. Percentages of agreement between FA and CA across the seven clusters ranged from 71% to 100%. The comparison of FA with MDS produced percentages of agreement ranging from 83% to 100%, and the percentages of agreement between CA and MDS ranged from 60% to 100%.

Table 2

Number of Common Jobs Found in Each Pair of Analysis Methods

	Cluster						
	1	2	3	4	5	6	7
<u>Method Pair</u>							
FA vs CA	20	18	19	10	13	9	9
FA vs MDS	24	18	17	12	13	10	9
CA vs MDS	18	17	20	9	9	8	10

As was mentioned above, the MDS solution produced 5 tightly grouped clusters and two more disparate clusters. The two disparate clusters were Cluster One and Three and of those two the more disparate was Cluster Three. MDS Cluster Three could be viewed as a combination of one or two smaller clusters and outliers.

From the MDS plot of the 113 jobs in three-dimensional space, it would appear that the large number of jobs in the CA Cluster Three was due to the forcing of all outliers into that cluster. When the differences in the FA and CA solutions were interpreted in the light of the MDS three-dimensional plots, the FA solution's disposition of the outliers was almost always supported more strongly by the MDS evidence than was that of the CA analysis. The FA tended to combine the outliers with clusters to which they were in more close proximity in three-dimensional space while the CA tended to lump them all together in one large cluster.

DISCUSSION

These analyses were exploratory in nature and any results determining the number of clusters are highly tentative. The focus of this study was on the comparison of the three methodologies. The three methodologies provided similar job groupings with some variations; the agreement across the three methodologies was 83%. The variations appear to be due to the treatment of outliers. While each methodology gives different kinds of information, multidimensional scaling provided the best information about the outliers.

Cluster analysis gives clean clusters. Each observation starts out as a cluster by itself and then the two closest clusters are merged to form a new cluster replacing the two old clusters. The merging of clusters continues until there is only one cluster. As a result, deciding how many job groupings to use becomes a subjective process. We chose seven job groupings which had an R-squared accounting for 70% of the variance.

Factor analysis provides more information than the cluster analysis procedure. In addition to the primary factor loadings, loadings on other factors (clusters) are given. It was found that often the sizes of the primary and secondary loadings were very similar, meaning that an occupation could go in one group as well as another, or that the occupation was an outlier. Gandy (1979) also concluded that Q-factor results were more interpretable than those from hierarchical cluster analysis when he grouped jobs for validity generalization purposes.

Multidimensional scaling gives a graphic picture of the interrelationship among the different groupings. The three-dimensional representation of the interrelationship among the occupations facilitated the placement of outliers.

Each of the three procedures provided different types of information. We agree with Gandy (1979) that factor analysis provides more interpretable results and recommend using MDS in combination with factor analysis because of the insight into the structure of the data which its data plots provide. Harvey

(1986) cautions against the use of MDS because of the high computer costs, but it was found not to be any more expensive than the other procedures, especially when ties were broken beforehand. Subjectivity is inherent in all three procedures especially when naming, or identifying, factors, clusters or dimensions.

REFERENCES

- Andrews, T.G., & Ray, W.S. (1957). Multidimensional psychophysics: A method for perceptual analysis. Journal of Psychology, 44, 133-144.
- Barnes, M. & O'Neill, B. (1978). Empirical analysis of selection test needs for 10 occupational groups in the Canadian Public Service. Paper presented to the meeting of the meeting of the Canadian Psychological Association, Ottawa, June, 1978.
- Christal, R.E. (1974). The United States Air Force Occupational Research Project (Technical report AFHRL-TR-73-75). Brooks Air Force Base, Texas: Air Force Human Resources Laboratory.
- Cornelius, E.T., Carron, T.J. & Collins, M.N. (1979). Job analysis models and job classification. Personnel Psychology, 32, 693-708.
- Cornelius, E.T., Schmidt, F.L. & Carron, T.J. (1984). Job classification approaches and the implementation of validity generalization results. Personnel Psychology, 37, 247-260.
- Fleishman, F.A. & Quaintance, M.A. (1984). Taxonomies of human performance. Orlando, Fla.: Academic Press.
- Gandy, J. (1979). Cluster analysis versus factor analysis in defining job groups. Presentation at the Conference of the Military Testing Association, San Diego, CA.
- Harvey, R.J. (1986). Quantitative approaches to job classification: A review and critique. Personnel Psychology, 39, 267-289.
- McCormick, E.J. & Jeanneret, P.R. (1988). Position Analysis Questionnaire (PAQ). In S. Gael (Ed.), The job analysis handbook for business, industry and government. New York: John Wiley and Sons, Inc.
- Rosse, R.L., Borman, W.C., Campbell, C.H. & Osburn, W.C. (1984). Grouping Army occupational specialties by judged similarity. Unpublished paper, 1984.
- U.S. Civil Service Commission. (1969). Handbook of occupational groups and series of classes. Washington, DC: U.S. Civil Service Commission.

OCCUPATIONAL ANALYSIS: PRESENT AND FUTURE

Dr. Hendrick W. Ruck
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

The Occupational Analysis Programs currently operating in the U.S. Air Force has a rich and deep research and development (R&D) foundation. The Air Force has been proud to lead the R&D of job and occupational analysis technologies. This panel presents two visions of occupational analysis in the Air Force. First, Mr. Joseph S. Tartell, Chief, Occupational Analysis Division, describes the procedures, projects, and scope of the present operational Occupational Analysis Program. Following that, Dr. Hendrick W. Ruck, Technical Advisor of the Training Systems Division, describes and discusses technologies that have potential for affecting the way occupational analysis could be performed in the future. Following the two presentations, Mr. Tartell and Dr. Ruck discussed the needs and impetus for changes to the present analysis system and the problems associated with changing a successful system.

OCCUPATIONAL ANALYSIS: THE PRESENT

J S Tartell
Occupational Analysis Division
USAF Occupational Measurement Center

The United States Air Force Occupational Analysis Program consists of a relatively small band concerned with the definition of the jobs and tasks performed by individuals in the pursuit of mission accomplishment. In the author's view, this flexible, versatile application of research has fundamentally changed the personnel, classification, and training decision-making approach in the US Air Force.

There are a number of reasons this introductory statement is so strong. However, the primary reason the credibility of the Air Force Occupational Analysis Program is strong and the data it provides are valuable to decision-making entities lies in the fact that the mission is to collect and analyze data, not to make recommendations or decisions. Members of the Air Force functional, personnel, and training communities make the decisions. In the present context of Air Force decisionmaking, occupational analysis results are used to define what personnel are doing, to support personnel-related decisions, and to defend the need for training programs. An area that the USAF Occupational Analysis Program has diligently worked to stay away from is the evaluative function of how well the training system prepares the individual or how well the individual performs that job. Occupational analysis data are used in the accomplishment of a variety of evaluative functions with two provisos: first, the data serve as the source of what job incumbents do, not how well the incumbents perform; and second, the evaluative functions are accomplished by those responsible for evaluations, not by the data collection and analysis organization.

Before reviewing the type of data provided by the USAF Occupational Analysis Program, there is one additional area to be discussed. Up to this point in this presentation, most comments have addressed the providing and use of DATA. But, that is not the most important product or output of the Occupational Analysis Program. The most important product is INFORMATION. As an organization, data are vital for the accomplishment of our mission; data is the fuel to make the engine run. But, as with almost any engine, the function or mission is not to make the engine run. The mission is to get somewhere. The mission of the Occupational Analysis Program is to provide fuel for the decision-making process. The primary reason for the existence of the Occupational Analysis Program is to furnish objective information in a clearly understood manner to managers and executives responsible for personnel management decisions.

To aid decisionmakers in performing their functions, the Occupational Analysis Program provides a variety of data in a number of different formats. First, what kinds of data and then the different formats. The USAF Occupational Analysis Program foundation is the percentage of job incumbents who perform a task. This single item of information has been and continues to be the keystone to the personnel management process--what is done and by how

ary incumbents. The utility of this single item is reasonably clear--its applicability to classification, training, testing, and other functions has been the subject of numerous presentations. To provide additional information, the Occupational Analysis Program uses Task Learning Difficulty and Training Analysis data. These are referred to as secondary factors, primarily because they exist in addition to the primary data and are used for a specific purpose. That purpose relates to training applications of occupational analysis information, while the primary factor (percent performing) has more than one application.

These are the normal types of data collected and reported from the Occupational Analysis Program. A key issue and one that is of prime value to the user is the translation of these data into information. There are two primary types of product which accomplish this translation--the Occupational Survey Report and the presentation of task data in user familiar formats. The first of these is a narrative report of the data which describes the jobs and tasks being performed, who is doing which tasks and what are the implications of the data relative to existing classification and training programs. The second product is a series of data extracts arrayed in a matching format--the occupational information is matched to the training documents (normally specialty training standards and plans of instruction). The value of this approach lies in the familiarity of users with the format--the information is arrayed in a manner and form that those who will use that data recognize and understand. There is a third part of this information presentation process which aids the user in applying the occupational information to a specific issue. This third part is the involvement of the occupational analyst with the decisionmakers as they interpret and apply the information to solve or resolve particular issues. A normal, routine, but critical aspect of every occupational analysis project is the personal delivery and presentation of findings to those concerned with the output. In every instance when an occupational analysis project is completed, managers and executives from the personnel, training, and functional communities are personally briefed on the findings. Additionally, further analysis may be done should questions or issues arise from these presentations which were not anticipated at project initiation.

It may be inferred from the previous paragraphs, a major focus of the occupational analysis process is the user. Over the history of the USAF Occupational Analysis Program, the schedule of projects has changed and the types of users have expanded. With the operational implementation of the Program, the primary users were the enlisted classification and training communities of the Air Force. As the Program grew and expanded, the training applications also grew. Use of occupational analysis information to determine relevance of training content continues to be a primary application. However, the use of occupational analysis information by functional management staffs has grown dramatically in the recent past. The expansion of responsibilities, coupled with attempts to reduce the numbers of specialties in the Air Force through such initiatives as RIVET WORKFORCE, have resulted in a need for comprehensive information regarding the tasks and jobs performed within those specialties. Additionally, with the shrinking number of specialties, there is an increased need for information concerning the types of jobs which exist within specialties in order to build career paths to prepare junior personnel for mid-level positions and then for management and executive responsibilities.

Not only have the uses of occupational information grown, the number and types of requestors for the information have expanded. When the program began, the aim was to survey each enlisted specialty once every four years. As the program's acceptance grew, special requests from the field also grew until today the USAF Occupational Analysis Program operates on a request basis--projects normally are not scheduled unless there is a specific user who has a definite need for the information. Over the history of the Program, the requestors traditionally have been the training managers (located at the six Air Training Command (ATC) technical training centers) or training staff officers (located at Headquarters ATC) who perceived a need to review the content of the existing training programs or needed information to build new training programs. More recent history shows functional managers (at the Air Staff or MAJCOM levels) requesting occupational analysis information to support or reject force structure plans. For example, a recently completed occupational survey of Intelligence Officers was requested by the force management staff within the AF functional management community, with a major purpose to review and compare the jobs and tasks performed by officers and enlisted personnel. Occupational information from the survey was used for the restructuring of some jobs such that enlisted personnel now perform the technical intelligence jobs while the officers perform leadership and management functions. Another recently completed project was requested by the Interservice Training Review Organization (ITRO) Space Training Task Force to determine the fundamental knowledges of the space environment necessary for nonoperators assigned to the joint space arena (this project was unusual not only because of the requestor, but also because of what was requested). Still another unusual request was that received from the civilian management of depot-level civilian aircraft mechanics--a project aimed at defining the requirements for job qualification standards and on the job training programs. Another different type of request involved the requirement to provide occupational information across the complete personnel spectrum of an occupational field--officers, enlisted personnel, and civilians. This approach allows the requestor to view in a single information source the complete area of responsibility. The ability to determine which segment of the workforce is truly performing which jobs and tasks and what are the areas of overlap among the different populations, allows a comprehensive review of functional responsibilities and a valid and reliable departure point for functional realignment of jobs and tasks.

And what of the success of the Occupational Analysis Program? Two examples will illustrate how functional and training managers made substantial changes based on occupational analysis information. The first example is the Aircraft Fuel Systems Mechanic career ladder where the occupational survey revealed large percentages of personnel working on external fuel tanks. In reviewing the career ladder classification description and training plans, the occupational analyst found no mention of external fuel tanks. The specialty was performing tasks for which there was no formal training, training that was provided was accomplished through OJT. Based on the findings from the occupational survey, the career field managers rewrote the classification description and then established a resident training program to support the needs of the operational community. This application serves to illustrate a number of different applications of the occupational analysis information--a change in classification documents, an addition to a resident training program and a decrease in the OJT responsibilities of the workforce.

A second example involved the Munitions Systems Maintenance career field where occupational survey information confirmed the removal of a responsibility from the field. Utilizing the data-matched plan of instruction, training managers were able to determine which specific blocks of instruction could be deleted. This example was particularly useful because the occupational survey information contained both electronics principles as well as task data. The reduction in the training program was approximately 30 days.

These two examples illustrate data-based decisions and reflect the primary reasons for the success of the Occupational Analysis Program: first, a valid and reliable program whose findings can be verified; and second the application of those findings is by the users not by the analysts.

The USMC Occupational Analysis Program has refused to become complacent despite a substantial number of successful years of operation. As the technology of weapons systems becomes more complex and the costs of personnel continue to escalate, the need for more information in different formats with greater reliability grows. One interesting change that has occurred in the requests for occupational information is illustrated by a request from the office responsible for pilot training within the Air Training Command. This request involves the application of occupational analysis techniques to determine the skills and knowledges required to pilot a variety of airplanes. This project entails the logical extension of the technology utilized in the Electronics Principles Inventory and accompanying analyses. While the task-based approach to job analysis provides a comprehensive basis for decision-making, a more in-depth level of information is required for entry-level training programs which demand an understanding of fundamentals prior to performance training.

One issue arising from the request for fundamentals types of projects (those involving skills, knowledge or other than task data) relates to the need for extensive objective information to support task analysis which, in turn, supports training development. The long-term, yet cloudy, distinction between education and training continues as an area of discussion when training programs are designed. With the increase in the requirements for more information at greater levels of detail (beyond the task level) the number of purely practical problems increase. The goal, at one time, was to complete an occupational analysis which would identify the tasks within an occupation and then perform a complete task analysis to determine all of the underlying elements required for designing training programs for all tasks. Practically speaking this is not possible. There are too many tasks in occupations and too few personnel available for training design.

As a partial solution to this dilemma, the task clustering concept allows the collapsing of long lists of tasks into a much smaller number of clusters. These task clusters have the advantage of being based in the occupational data which means that the tasks have some basis for being grouped together. That reason must be determined through analytical procedures; but, regardless of the reason, there is an objective basis for training the tasks together. The logical extension of task clusters is cluster analysis for underlying constructs. This approach lessens the workload without diminishing the

utility of the information. Additionally, by using the occupational information which indicates which segment of the occupation performs the tasks within any cluster, and considering other data such as the equipment used or job location, training designers, developers and managers will be able to specifically target training to those who will benefit the most.

Training design and development is one area where practical considerations must drive the application of technology to meet training needs. Within the Air Force, the Training Development Services Division of the Occupational Measurement Center is working on practical solutions to design issues. The preparation of Training Requirements Analysis Reports provides a partial answer by reporting the results of specialty-wide task analysis and the need for career training programs. Present approaches are limited by practical concerns, such as the sheer numbers of tasks and the volume of information necessary to support task analysis. Projects utilizing the clustering technique are underway and results will be reported.

Another area of change in the occupational analysis world involves the awareness of and ability to include information from other data sources. There are a large number of databases which contain relevant information. Both the logistics and maintenance communities have large scale databases with information useful to training personnel. Conversely, personnel-based occupational information will be useful to the logistics and maintenance communities. On top of these is the weapons systems acquisition community which requires information from all sources to design and plan for the implementation of new systems. This need has been recognized. The CROSSWALK and FOOTPRINT projects are attempts to bridge the existing gaps by linking equipment and specialty databases.

The Air Force Occupational Analysis Program's recognition of some of these new concerns and potential sources of information, coupled with efforts to incorporate the latest technology into our methodologies, confirms our resolve to continue to grow and respond to the ever-changing needs of the user community. This forward-looking approach, combined with 20 years of research and operational experience, a willingness to continue to collect and analyze occupational data, and a refusal to pontificate from survey results will ensure the Air Force Occupational Analysis Program continues to be an integral and respected cog in the machinery of Air Force personnel management, classification, and training programs well into the 21st century.

OCCUPATIONAL ANALYSIS: THE FUTURE

Dr Hendrick W. Ruck
Training Systems Division
AF Human Resources Laboratory

The Armed Services have been using occupational analysis (OA) technologies operationally for more than 20 years. Research has been continuing in this area for an even longer period of time. This paper will address what occupational analysis could look like in the next century. Further, the paper will explore the reasons leading to the changes that are being proposed and will suggest necessary research and development (R&D) activities.* Christal and his associates discussed the impact of the computer in occupational analysis and personnel management in the early 1960s. They pointed out that computers had shown that rapid calculation was then a possibility as of that time. Problems that would take months to solve could be solved in hours with the use of computers. However, they noted, the primary role of the computer was in science and engineering. Their R&D was one of several initiatives that brought computers to personnel management. Over the course of the past 27 years, computer science and computer engineering have advanced so dramatically that consideration should now be given to how these advances could effect occupational analysis in the 21st century.

Historical Context

In order to establish a perspective for the future, a short review of the history of the development of the occupational analysis technology is appropriate. In the early 1960s, the R&D of occupational analysis methodologies centered on the collection of survey data; on the construction and behavior of scales such as time spent; on the use of computers for cluster analysis of jobs; and on the use of occupational data in job evaluation. It was not until the late 1960s and early 1970s that R&D on the task difficulty scale and on a composite index (average task difficulty per unit time spent, ATDPUTS) was performed. The primary goal of occupational analysis research

focused on the collection and analysis of data to support the personnel classification system in the military. Although personnel applications were a focus, Christal noted that one of the major uses for occupational data could be the development, evaluation, and revision of course curricula and training requirements. During the fledgling years of the Air Force OA Program, trainers found the data to be quite useful, but somewhat lacking (given all of the published guidance on instructional system design (ISD)). In particular, ISD variables were often not addressed in the OA program. Hence, R&D into instructionally-related scales (task delay tolerance, possible consequences of inadequate performance, recommended training emphasis and hazard potential) was performed through the 1970s. The R&D performed during the 1960s and 1970s resulted in an effective and efficient system for collecting, managing, analyzing, and reporting occupational data to support both the personnel and the technical training systems in the Air Force.

Why Change?

Given that the Air Force has an R&D based, operationally proven OA system, one might wonder why I would propose changes. There are two different forces acting on the OA methodologies that suggest change is needed: changes in the Air Force, and significant advances in computer technologies. There is a need to perform R&D to develop revolutionary new approaches to performing occupational analysis. What are the pressures in the Air Force? The most important change in the Air Force personnel system is the reduction of the number of Air Force specialties. This is especially the case in the maintenance area where the boundaries between specialties are deliberately being blurred to provide managers with the opportunity to utilize their personnel in a variety of tasks and jobs that are not presently contained within a single specialty. This blurring will necessitate longer task lists than have ever been used before in the OA process. A second pressure is the strong emphasis being placed on the measurement of jobs using tasks directly linked to weapon systems. Thirdly, the present occupational analysis system, while effective and efficient, has been tasked as much as possible; there is little, if any, "surge capacity."

In addition to these pressures on the present system, there are a number of recent technological advances that offer the opportunity for changing the current approach to performing OA. These technologies include (a) the fielding of a disbursed-computer based personnel-management system; (b) the emergence of computer-based management systems for different functional areas within the Air Force; (c) the maturity of computer adaptive testing as a science; (d) the development of significant advances in expert systems technology, and artificial intelligence, particularly as they relate to database management and linguistic analysis. The remainder of this paper will discuss how these technologies could affect data collection procedures, inventory development, data management, analysis, and reporting.

Data Collection Procedures

In thinking about using computers rather than survey booklets to collect occupational analysis data, a number of different options are possible. These include the application of latent-trait theory, the development of computer-based delivery systems, and the exploration of the potential for intelligent occupational surveying.

First, consider the application of latent trait theory to occupational analysis. Because occupational analysis programs use comprehensive data bases full of tasks and associated probabilities of performance, one could construct a system which would present tasks to respondents based on latent trait theory principles. That is to say that only selected tasks would be presented to respondents based on known statistical relationships of the data associated with tasks. This would have the advantage of presenting on a limited list of tasks from the existing data base to each respondent based on the probability that the tasks are performed by the individual. Research in this area would focus on determining the stability of the probabilistic structures underlying the OA task data base; in addition, research would have to be performed to determine how to update the relationships within the database.

A second approach would involve using the computer to "dumbly" capture responses and to simply administer job inventories on computer screens.

While there are considerable advantages related to ease of data capture with this approach, there are some research questions that would have to be answered regarding the quality of the scaling and the impact of computer-based surveys on response reliability and validity.

A third approach would be to use artificial intelligence procedures in the development of an "individualized" task list in real time as the respondent interfaces with the computer. The computer would query the respondent with a series of questions regarding his/her job. Artificial intelligence techniques would then be used to develop a set of tasks that are relevant for this individual. The individualized set of tasks could then be presented for scaling. Clearly, a number of research issues are raised in such an approach.

Using any of the above described approaches would have a number of advantages in the occupational analysis program. First, capturing individual responses via computer would reduce administration requirements significantly; second, data could be captured on a continuous (for example, once every two years per individual) basis. One might consider having individuals respond during their personnel record reviews, for instance. Third, if continuous updating of the data base were performed, current research questions regarding the quality and utility of "old" data and the rapid changes in the world of work could be researched.

Data Management

Present data management in occupational analysis is performed one study at a time. Task and response data are typically stored and analyzed as a unit. There is little or no cross-feed among occupational analysis studies. With advances in artificial intelligence approaches to data management, cross specialty data sharing could become an important capability for analyzing the current occupational structure and highlighting potential structure changes based on data commonalities. To build such systems would require a considerable amount of research. This research would include correlations of task statements using linguistic analysis necessary to support translation of tasks among specialties and to interpret tasks which appear to be the same in

different specialties. This is a high risk area; however, given advances that are being made in linguistic R&D, this could have very high payoff. Technologies such as hypertext and other artificial intelligence data base management approaches would have to be explored to build systems to meet their need.

Analyses

Occupational analysis is difficult, laborious, and time consuming. Occupational analysts spend considerable amounts of time poring over products to understand the "true structure" of occupations. A considerable amount of work is performed at the front end of each analysis to ensure that all of the computer runs to perform difficult and complex analyses are available in a timely fashion. Using expert system approaches, this front-end work could easily be streamlined. Further, tentative analyses of the structures using OA analyst expert systems could be made. Such expert systems are possible to develop because of the quantity and quality of OA expertise available. These expert-system-based analyses would require human reanalysis for verification and validation. The human's task would be to understand the analysis and fine tune it, rather than to develop it starting with almost no initial information. The human in such a system would be able to spend his/her time and energy on interpretation, special applications, and on significant management issues. Research in this area would have to first develop systems to capture analyst expertise regarding the OA processes. There is a certain amount of risk in this approach. Specifically, researchers have not carefully studied occupational analysis to determine the similarity of different analysts' approaches. Additionally, there has been little research to examine how similar different analyst's results would be. Research into expert systems for analysis would quickly develop answers to the first question, and quite possibly, the second one.

Report

Reports typically produced from OA organizations are usually written to cover a standard range of issues with exceptions made for known special interest

items. Often, such reports are less than optimal in terms of communicating to the reader the information he/she needs. Again, using advances in computer based technology one can imagine the development of a series of standardized reports targeted for different users. The human input would be on the analysis and discussion portions of the report and on tailoring the report to specific users. This would enable two, three, or more reports to be generated for each study, each for a specific user.

Summary

R&D of occupational analysis technologies and the trends in computer-based technologies outlined in this paper have the potential for radically changing future occupational analysis systems. Occupational analysis in the 21st century could use task inventories developed using existing databases through expert systems. Further, administration of inventories could be largely automated through personnel based computer systems or functional computer systems. Data base management could be improved by reducing the boundaries between specialties and studies in the database. Finally, analysis and reporting could be artificial-intelligence-based to allow the human analyst to focus on interpretation, special applications and important management issues.

DEVELOPMENT AND DESIGN OF THE NAVY OFFICER OCCUPATIONAL DATABASE

Captain Edward L. Naro, USN, and Dr. Janet M. Treichel
Navy Occupational Development and Analysis Center
Washington, D. C.

During the past three years, the Navy Occupational Development and Analysis Center (NODAC) has been developing an officer occupational database using a two-phase approach. Phase I has involved the administration of a generic managerial instrument, the Officer Survey Instrument (OSI). This instrument was developed by modifying the Professional Managerial Position Questionnaire (PMPQ) by Drs. Ernest McCormick and Jimmy Mitchell. The purpose of the OSI is to gather baseline data on common managerial functions and general position processes that cross all Naval officer designators and communities. Phase II consists of community-specific surveys. These instruments are task-based occupational surveys that employ the Navy Occupational Task Analysis Program (NOTAP) methodology and provide data on current tasks being performed by officers in specific communities.

Together, the OSI and the community-specific surveys will provide data to Navy manpower, personnel and training (MPT) decision makers that can be used to improve the use, training and assignment of personnel. This panel presents an overview of the Navy officer occupational database and the development, implementation and application of the OSI and the community-specific surveys. The discussions included the specific application of data obtained from NODAC's initial officer community-specific survey--the Medical Community.

**DEVELOPMENT AND DESIGN OF NAVY'S
OFFICER OCCUPATIONAL TASK ANALYSIS PROGRAM**

Captain Edward L. Naro, USN
Officer In Charge,
Naval Military Personnel Command Detachment,
Navy Occupational Development and Analysis Center,
Washington, DC

Over the past 3 years, Navy's Occupational Development and Analysis Center (NODAC) has been engaged in developing and implementing a new officer occupational analysis program. This program, which includes conducting officer occupational task surveys, creating an officer occupational data base, and performing occupational analysis using that data, is intended to provide information needed to support decision making throughout Navy's Manpower, Personnel, and Training (MPT) arena.

Background:

In the days of "wooden ships and iron men" when the US Navy first came into being, the classification of its officer structure was relatively uncomplicated. There were officers of the line, ... Admirals, Commodores, Captains, Master Commandants, and Lieutenants; and there were Surgeons, Surgeons Mates, and Chaplains. In addition there were seven types of Warrant Officer to cover "technical specialties" such as carpentry, gunnery, and sailmaking. Less than 20 labels were used to identify officer requirements and the officers who possessed them.

Today Navy's officer structure includes 14 Officer "Pay Grades" which reflect rank; and over 140 "Designators" which identify careers, warfare specialties, staff corps specialization, regular or reserve status, and eligibility for, or limitations on, assignment to positions of command. In addition to these primary classification labels, there are now over 255 "Subspecialties" used to identify specific technical knowledges and skills obtained through either formal post graduate education or experience on the job. There are also 896 "Additional Qualification Designators" used to identify specific qualifications, i.e., experience in a particular type aircraft or weapon system. Finally there are 741 "Officer Billet Classification Codes" used to describe what the incumbent of a billet does.

The increase in complexity of the Navy's Officer Corps structure has been, for the most part, concentrated in the last century, with the most dramatic increases occurring over the last 50 years. This is primarily attributable to specialization resulting from technological growth that provided new platforms and weapons systems and made warfare at sea multidimensional, i.e., surface, subsurface, amphibious, air, and space. Research and development efforts now underway on highly sophisticated weapons systems, platforms, and communications/information systems

planned for the 21st Century promise even more rapid technological growth and continued pressures toward specialization of Naval Officers.

Opposing the trend toward specialization within the Navy's Officer Corps structure is the requirement for the Naval Officer to be somewhat of a "generalist" ... a traditional concept strongly reinforced by: (1) success in past conflicts, (2) the need to "grow" senior officers who are versed and capable of leading across many specialties or in more than one dimension of war at sea, and (3) physical limitations associated with manning of today's ships.

Balancing these opposing requirements ("generalist" vs "specialist") so as to achieve and maintain the optimum mix over the next 20 to 30 years promises to be a major challenge. To provide Navy's leadership and MPT managers with occupational analysis and information needed on a continuing basis to meet that challenge, the Occupational Development and Analysis Center has begun a long range Navy Officer Occupational Task Analysis Program (Officer NOTAP).

This new Officer NOTAP program differs significantly from past approaches to officer occupational analysis in at least two important respects. First, in the past, officer occupational survey and analysis efforts were only done on specific communities and only when requested. Such efforts were for the most part sporadic and reactive. A community had to recognize a problem and request a study before an attempt would be made to even gather data needed to conduct an analysis. As a result, communities requesting a study had to wait two years or more for results. The Officer NOTAP program now being implemented calls for surveying and conducting an analysis of each community, addressing issues which surface in the analysis as well as those already known to the community, and maintaining the survey data in an occupational database for ready use and comparison with subsequent survey data.

The second difference between the newly established Officer NOTAP and past approaches to officer occupational analysis is a function of scope. Past officer occupational analysis efforts by NODAC have been unique with each limited to a specific concern. As a result, comparing occupational data between communities was not possible. To be sure, the new Officer NOTAP program will include community specific surveys, however, these surveys are being designed to allow compatibility of data across community lines. This will permit occupational analysis to address concerns and issues common to more than one community. Examples of past changes to the structure of the Officer Corps where such comparisons would have been extremely beneficial include the establishment of the Weapons Systems Acquisition Manager (WSAM) program, and later the establishment of the Material Professional (MP) program. Other areas where occupational analysis across community lines will be useful include: the use of Line Officers in the Medical Department; LDO and Warrant Officer programs; Task cross over

between officers and enlisted members in the combat information arena; and Naval Officer educational and training requirements in general.

Program Development:

Establishing the Officer NOTAP program began with initial discussions between NODAC staff members and the MPT managers and Warfare/Resource Sponsors. These discussions and related briefings ensured that potential users of the occupational information would be involved from the start, that they would understand and support the survey and analysis program, and that the information produced would satisfy their needs. These initial discussions and briefings also set the stage for more detailed discussions that would be held later during the research phase of each community specific survey.

The overall survey plan calls for using two types of survey instruments. The first, a single, general managerial and professional responsibilities survey called the OFFICER SURVEY INSTRUMENT (OSI), is being used to survey a sample of the entire Officer Corps. It has been administered to all communities and grades from Warrant to Captain. The OSI is derived from the Professional and Managerial Position Questionnaire (PMPQ), which was developed by Drs. Jimmy Mitchell and Ernest McCormick. Upon obtaining the PMPQ from Purdue University, NODAC adapted it for Navy application. It provides data on general managerial and professional activities (financial management, communications, etc.) that we expect may be found in varying degrees in any officer community and at any grade level.

The second type instrument is a community specific task survey, tailored for each officer community. These surveys exclude items covered in the OSI and focus on the other tasks being performed in the surveyed community.

This dual survey approach will permit comparisons on general managerial and professional responsibilities between pay grades and between communities and designators, while at the same time permitting a more detailed analysis of tasks performed within each community. Although administered separately they are intended to complement each other. Combining the data from both types of survey will provide the comprehensive occupational database that is needed to support present and future decisions affecting the structure of Navy's Officer Corps.

Planning and Scheduling:

The first OSI was scheduled to be conducted and completed by the end of CY 1988. An analysis of the OSI data is to be conducted early in CY 1989. Data obtained from the OSI will also be combined with data obtained in each community specific survey so that the analysis of each community will include all data available on that community. The OSI, or a modified version of it

will be administered again before the next cycle of community specific surveys is begun.

In planning and scheduling the community specific surveys, we divided the Officer Corps into four major components: Aviation; Surface; Subsurface; and Staff. Each component was then further subdivided into "communities" to be surveyed. For example, Staff was divided into Medical, Legal, Chaplain, Civil Engineering, etc.; and Aviation was divided into Anti-Submarine Warfare (ASW), Tactical, Support, Maintenance, and Other Aviation. Completing the entire schedule of surveys and analysis is expected to take a minimum of ten years, at which time the survey cycle will be repeated. This timeline is based upon using existing personnel and resources. Availability of additional resources would allow the first cycle to be shortened considerably. In any case, future survey cycles should be shortened by our progression on the learning curve.

The Medical community was the first community scheduled to participate in a community specific survey. Aviation communities are scheduled to follow Medical. A preliminary schedule for communities following the Aviation communities place Surface next followed by Subsurface, and then the remainder of the Staff Corps. This schedule may be altered by changes in priority dictated by MPT managers or Warfare and/or resource Sponsors.

Survey Methodology:

Methodology for conducting the Officer NOTAP surveys (both the OSI and the Community Specific Surveys) is similar to that used for many years by NODAC in the conduct of its Enlisted NOTAP surveys. Each survey includes the following phases: (1) Research; (2) Observation and Interview; (3) Task Inventory Development; (4) Survey; and (5) Data Entry.

RESEARCH: During this phase, project officers conduct research on any relevant past studies, the classification code structure, the billet file, personnel inventory information, etc. They also visit and coordinate with MPT managers, Warfare and/or Resource Sponsors, NODAC analysts, and subject matter experts (SME's) to ensure that the survey will cover all important issues and obtain needed data. Based upon this research, the population to be surveyed is defined, demographic questions are developed, a strawman task inventory is created, and a plan for conducting observation and interview (O&I) visits is developed.

OBSERVATION AND INTERVIEW: Teams of officers visit fleet and shore activities on both coasts (and overseas when appropriate) to observe and interview officers in the field, in the community to be surveyed. During these visits new tasks are added to the inventory, existing tasks are modified as appropriate and fleet issues which need to be addressed are highlighted in either the task statements or demographic questions.

TASK INVENTORY DEVELOPMENT: Upon returning from the O&I visits, the team members combine their task lists into a master task inventory. Project officers then consolidate and combine task statements where appropriate to keep the survey instrument within practical limits while ensuring sufficient specificity to address issues and concerns raised in the research and O&I phases. The task inventory and demographic questions are then finally reviewed by MPT managers and Warfare and/or Resource Sponsors before being incorporated into a survey book. As a last step in this phase, if the survey book required significant design changes, it is pre-tested prior to final printing.

SURVEY: Sample size is determined by the project officer and NODAC analysts using a matrix of cells describing combinations of grade, designator, type command, etc. The surveys are mailed out in command bundles with each survey addressed to an individual officer. Mailout and responses are tracked to ensure required numbers, both overall and in each cell, are received. Where insufficient returns are received, we follow up with tracers.

DATA ENTRY: Upon receipt, surveys are logged in, checked for completeness and scanned into the mainframe and stored on magnetic tape. The data is processed using CODAP and SPSS-X.

Analysis:

At the conclusion of each survey, we will conduct an analysis of the data and provide reports and briefings on that analysis with recommendations to the MPT managers, and Warfare and/or Resource Sponsors. Analysis addressing issues that cross community lines will be scheduled as data becomes available. In addition, NODAC will perform analysis and provide data for users on request.

Progress to date:

The OSI - The OSI survey, I am pleased to report, is nearing completion. We expect to begin analysis in January 89. To provide more details on that project, Lieutenant Sue Fiorino, the OSI Project Manager and a recent graduate of the Navy's Post Graduate School MPT Curriculum where she earned a Masters Degree in Personnel Management, will be presenting a paper appropriately entitled **The Officer Survey Instrument.**

Community Specific Surveys - The prototype for the community specific surveys is the Medical community survey. It has been completed and is now in the final stages of analysis. Reports to the Medical Command and to the Navy's Surgeon General will be made by the end of the first quarter of CY 1989. A mainstay in this effort has been LCDR Ellen Quisenberry, Nurse Corps, who is now assigned to Navy's Medical Command. LCDR Quisenberry, who also earned her Masters Degree in Personnel Management at the Navy's Post Graduate School, was assigned to NODAC for eighteen months where she was project manager for the Medical Community Survey.

She will be following Lieutenant Fiorino, presenting a paper entitled The Navy Medical Community Officer Occupational Task Analysis Program (NOTAP) Survey. While her paper focuses primarily on the Medical Community Survey, it is important to note that it is presented here as a detailed example of our community specific surveys. Many of the lessons learned in that survey are being applied to the next community specific survey (the Aviation ASW community survey) which is now in the task inventory development phase.

THE OFFICER SURVEY INSTRUMENT (OSI)

LT Susan J. Fiorino, USN
Navy Occupational Development and Analysis Center
Washington, DC

The U. S. Navy's officer structure has evolved over time in response to changes in the complexity of war at sea and resulting growing requirements for new skills, experience, education and training. The speed with which these requirements are changing, and the piecemeal changes to the structure that have resulted, suggest it is time to review the officer structure as a whole to determine whether or not it is optimum for present requirements, and to establish an occupational baseline reference to support an orderly transition of the Navy's officer corps into the 21st century.

In order to begin reviewing the officer structure, and the training, manpower and personnel policies and classification system that support it, an occupational database must first be established. This is being accomplished at Navy's Occupational Development and Analysis Center (NODAC) through a long-range and continuing program of occupational surveys and task analysis. In addition to developing community-specific task inventories to survey the specialized tasks performed within the various officer communities, NODAC has developed a general managerial Officer Survey Instrument (OSI) to obtain data on the common managerial and professional functions which one would expect to find in varying degrees in any officer community and at any officer paygrade. Administering the general OSI to complement the community-specific surveys permits greater emphasis on community-specific tasks in the community-specific surveys while providing data required to compare like functions across community lines. This paper will be used to describe the development and administration of the OSI survey, and projected analysis of the occupational data it provides.

The Professional and Managerial Position Questionnaire (PMPQ)

When the decision was made in 1986 to administer a general managerial survey, two options were available for building it: (1) NODAC could develop an entirely new survey instrument, or (2) an off-the-shelf instrument could be adapted for the Navy's use. Due to the extensive time and cost required to build and validate a completely new survey, the second option was chosen.

The survey instrument chosen for adaptation was the Professional and Managerial Position Questionnaire (PMPQ), developed in 1976 at Purdue University by Drs. Jimmy Mitchell and Ernest McCormick. The PMPQ was designed, and is used today, for describing and analyzing professional, executive and managerial positions. It has been used in a variety of industries across the country, primarily for pay equity issues for upper management positions.

As described in the introduction to the PMPQ, the questionnaire consists of three sections. The first section is structured to

elicit two responses related to specific job activities: (1) the degree to which each activity is a "part-of-the-job", and (2) the activity's level of complexity. A nine-point scale is used for each rating. The scale for the "complexity" portion also includes specific examples, which act as anchors for the relative points of the scale.

The second section of the PMPQ involves describing the personal requirements, such as education and training, which the incumbent believes to be important to adequate performance in the job. The third section contains miscellaneous information about the job, such as the requirement for licensing or certification, and the number of people supervised.

Modifications to the PMPQ

Before making any modifications to the PMPQ, we spoke with numerous individuals in the U.S. and Canadian military services, and with knowledgeable occupational analysts outside the military, to collect information and ideas. As a result, several modifications were made to the PMPQ to adapt it to the Navy's unique characteristics. The most significant modification involved changing the anchors on the complexity scales into examples which relate to Navy functions. Additionally, the salary information was deleted, as this can be obtained from the background data on grade and years of service.

In addition to modifying the information in the PMPQ, the questionnaire was also expanded to include job responsibilities important to Navy officers. These included questions regarding contract administration, pre-deployment planning, interservice interaction, application of military law, inspections, watchstanding, physical fitness and leadership. Additions were also made to the background items to include military-specific job and individual data, such as the number of years an individual has served as a commissioned officer, any service colleges attended, and the amount of time spent on collateral duties, watches and meetings.

The OSI

The net result of these modifications was a 37-page survey consisting of billet and personal background information, 33 questions regarding management and professional responsibilities, and four questions based on leadership functions. The management/professional items look very much like those in the PMPQ, with nine-point scales for the "part-of-job" and "complexity" portions, and anchors on the complexity scales.

Upon completion of the initial development phase of the OSI, a pre-test was conducted. The OSI was administered on-site in Mayport, Jacksonville, Norfolk, Portsmouth and Washington, DC, to 150 officers representing a variety of designators and grades. Additionally, the supervisors of these officers completed the complexity portion of each question, responding with respect to their subordinates' positions. This provided an additional check on the validity of the complexity variables.

In keeping with the findings of Perrin, et. al., which indicated that survey format affects individuals' responses, two forms of the OSI were pre-tested to determine which form better encouraged independence of judgment. The mean responses to each item were examined using t-tests, as were the correlations between parts A and B of each item. There was no significant difference in responses between the two forms.

Analysis of the pre-test data and discussion with the respondents revealed that some of the demographic items were being interpreted slightly differently by different individuals. The instructions were clarified and some items and anchors were re-worded to be more specific. The analysis also indicated that there was no significant difference between the complexity responses of the incumbent and his/her supervisor; therefore, we decided to administer the survey only to job incumbents.

After making modifications based on the pre-test analysis, a small second pre-test was conducted on 25 officers in the Washington, DC area. Results of this second pre-test confirmed that problems encountered in the first pre-test had been resolved satisfactorily.

Administration of the OSI

The eligible population for the survey was determined by examining the total officer population, as maintained in the Officer Master File. Officers in paygrades W-2 (warrant officer second) through O-6 (captain), either regular Navy or active Reserves, were eligible. Certain groups of officers were excluded from the eligible population because their responses would not be representative of the job information we were aiming to collect. For instance, officers filling billets as students were excluded.

The sample size was determined by first stratifying the eligible population by designator group (e.g., aviators, civil engineers, etc.), and then proportionating by paygrade. The survey mailout number was calculated by assuming a 65 percent return rate for each cell in the "paygrade by designator" matrix. A 65 percent return rate is assumed, based on past experience, to account for individuals who are not available, those who receive a survey but do not return it and surveys which are unscannable. This process produced a total mailout number of 10,897; the number of returns required is 7,127. Approximately 7,300 completed surveys have been received thus far. The survey remains open now as we confirm that we have sufficient returns in all individual cells.

Data Analysis

Analysis of the survey data is expected to begin in January 1989. Responses will be compared between and among designators, grades and job titles to determine how managerial functions differ among the various groups. This information will be used to review career progression within a designator, with the hypothesis being that the complexity of job functions performed increases with grade. Additionally, identification of the managerial functions

performed by officers in various designators, grades and jobs will help to identify the level and type of training or education required by officers in various careers and/or jobs. This information will also be provided for curriculum review in Navy training and accession programs.

The billet and personal information will be examined for consistency in assignments. For instance, incumbent designators will be compared with billet designators to see how closely assignments match stated billet requirements. A large percentage of mismatches could indicate there are problems either with assignments or with the statements of billet requirements. In either case, such a finding would invite further investigation by community managers and/or distribution managers. Job functions performed will be compared with subspecialty codes to determine the degree to which subspecialty skills are being utilized. This information will also be provided to education and training and distribution managers. Finally, job functions will be compared with billet codes to determine the accuracy of billet codes in describing the managerial functions associated with the billet. This information will be used to assist warfare resource sponsors and manpower planners in reviewing the coding structure used to describe the functions an incumbent is supposed to be performing in a given billet.

Summary

The OSI has been designed and is now being administered to the officer corps of the U.S. Navy to obtain data required to establish a comprehensive officer occupational data base. This data base will be maintained and periodically updated to assess the appropriateness of the existing structure and to support future planning. Together with community-specific surveys, the OSI will provide information to support needed changes in officer education, training, manpower planning, classification structures and personnel policies.

REFERENCES

- Mitchell, J. L. and McCormick, J. J. (1976). Professional and Managerial Position Questionnaire, West Lafayette, Indiana: Purdue Research Foundation.
- Perrin, B. M., Vaughan, D. S., Mitchell, J. L., Collins, D. L., and Ruck, H. W. (1987). Effects of Data Collection Format on Occupational Analysis Task Factor Ratings. Proceedings of the 29th Annual Conference of the Military Testing Association. Ottawa, Ontario, Canada.

THE NAVY MEDICAL COMMUNITY OFFICER OCCUPATIONAL TASK ANALYSIS PROGRAM (NOTAP) SURVEY

LCDR M. Ellen Duisenberry
Naval Medical Command
Washington, DC

The purpose of this paper is to present the prototype of NODAC's community-specific officer surveys. Together with the Officer Survey Instrument (OSI), these community-specific surveys will provide the basis for the Navy's officer occupational data base. The Navy Medical Department was the first community to be surveyed. Aviation and surface warfare communities are currently under study and all other communities are scheduled for survey in the out years.

The purpose of this community-specific survey was to assist with the analysis and description of professional positions of the Navy Medical Department Officer Corps. Broad uses of the data will include: (1) improved identification of manpower requirements, (2) improved description of personnel resources and (3) improved overall manpower, personnel and training management efficiency. Specific uses may include: (1) review of classification structure, (2) definition of grade progression and (3) definition of job similarities and differences by facility.

The project was initiated in August 1985 with a meeting between NODAC and Naval Medical Command (MEDCOM) task force representatives. Subsequent meetings were held between NODAC and MEDCOM to provide clarification regarding medical classification structures, special issues, or potential problems.

Information regarding job descriptions, tasks performed and associated job requirements (committees, collateral duties and watch duties), work group and resources, was collected by a team of NODAC personnel. This team conducted over 525 interviews worldwide in a variety of medical facilities including hospitals (CONUS and OCONUS), headquarters commands, medical and dental clinics, ships, Marine units, squadrons, Fleet Training Centers, Naval School of Health Sciences, Environmental Preventive Medicine Units and Health Research Center.

The survey instrument was developed to describe this complex community utilizing task statements and demographics. The vast amount of data collected during the interview process was reduced to pertinent task statements providing an accurate measure of what is occurring occupationally within Navy Medicine. Six duty categories provided the framework for the task statement section of the instrument. These categories were patient care; clinical support; dental, administrative and command functions; and miscellaneous. Through a process of ongoing coordination and consultation, the tasks were reviewed and refined to ensure completeness, accuracy and clarity. Demographic data was collected in three groups - information

about the billet the officer filled (e.g. billet designator and grade), information about the officer (e.g. SSN, grade, subspecialty codes) and information about the job (e.g. work area, job title, special duties).

A pretest of the instrument was conducted at two major sites, Southwest Region in San Diego and the Charleston/Beaufort, SC area. The pretest sites were chosen to provide a descriptive cross section of the medical community including both East and West coast locations.

The survey was mailed to 6931 officers in the fall of 1987. The Medical Officer community consists of approximately 12,000 officers in 9 paygrades and was distributed across 5 corps or designators). Our goal was a 95% confidence interval with 5% error. We expected a 50% response rate from the Medical Corps and 65% from all others. Response rates were based on statistical analysis and historical information and accounted for transfers, surveys not returned and unscannable surveys.

Analysis is being conducted on 4265 cases using both CODAP and SPSS-X. Due to the size of the survey and our computer limitations, CODAP runs were completed compliments of our friends at the Directorate of Military Occupational Structures in Ottawa. Comparison of respondent's task performance was completed using cluster analysis (CODAP OVLGRP and DIAGRM). As expected, Physicians, Dentists, Nurses, and Medical Service Corps (MSC) primarily clustered as separate groups. Subgroups were also identified within each major group. For example, Physician's Assistants (PA), with the highest degree of similarity, grouped with the Physicians. Clinical Psychologists (MSC) and Psychiatrists grouped as another physician subgroup. Nurse Anesthetist and Anesthesiologist grouped together as a subgroup within the Nurses cluster. Tab 1 demonstrates the initial cluster groups.

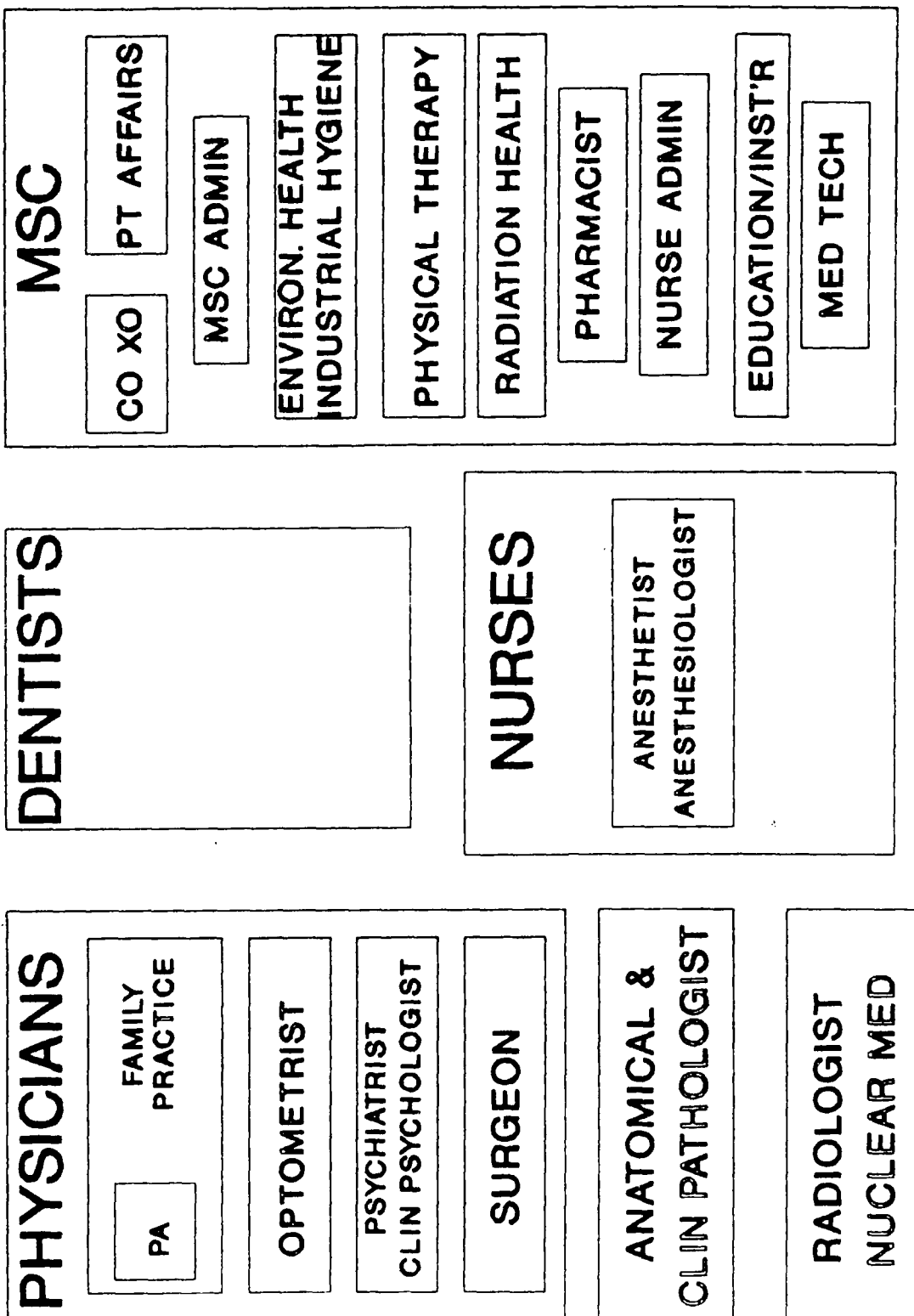
Distribution of cases for analysis is as follows:

21XX (Medical Corps)	994	23%
22XX (Dental Corps)	844	20%
23XX (Medical Service Corps)	1270	29%
29XX (Nurse Corps)	999	23%
75XX (Physician's Assistants)	160	3%

The 637 activities surveyed were categorized according to function resulting in 12 categories, including hospitals (5 sizes), surface, aviation, research, staff, dental (2) and field commands. Analysis will be conducted within each category as well as between categories. Additional break-outs will include distribution by corps and by grade for comparison between facilities and categories.

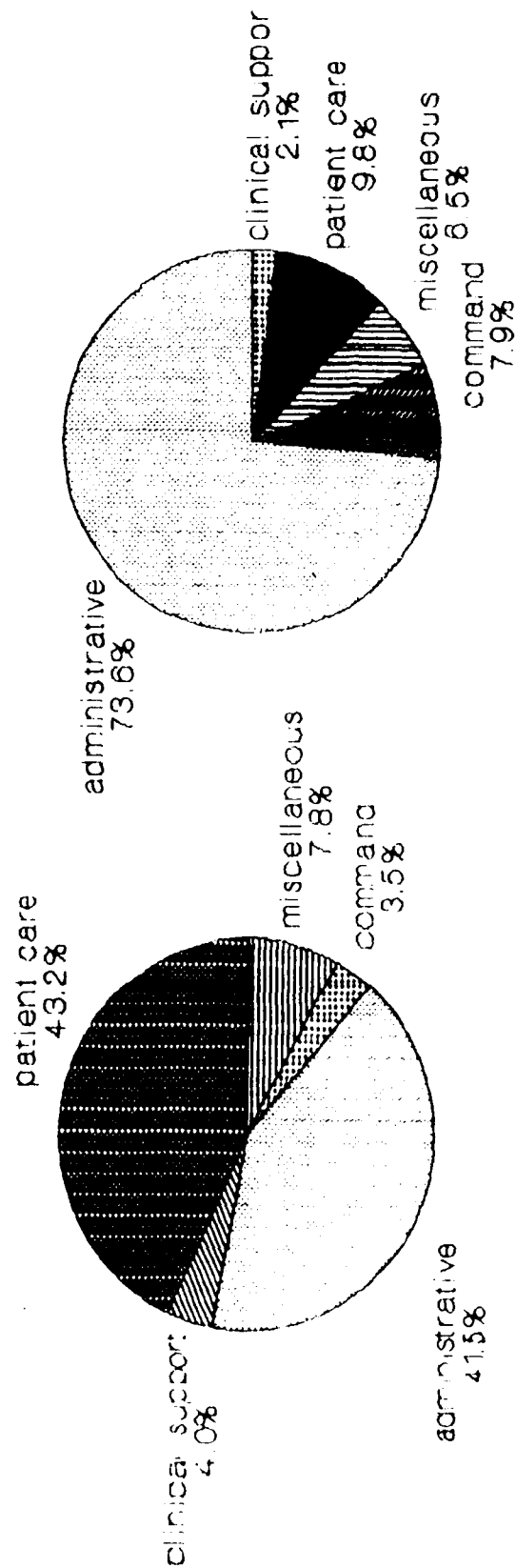
Tab 2 displays Nurse Corps Officers' task significance by duty category and hospital facility size. The first pie chart displays all Nurse Corps task significance in large hospitals. The second pie chart displays a specific group of nurses, the O5 or Commander group in large hospitals. This type of

CODAP GROUPS - NAVY MEDICAL COMMUNITY



Task Significance of all Nurses

Large Hospitals



All Nurses

Commanders (O-5)

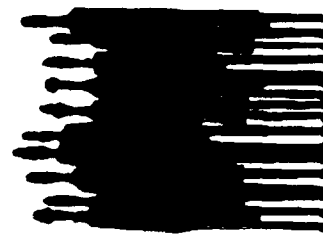
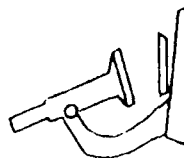
analysis provides the Nurse Corps with differences and similarities between facilities and paygrades allowing for description of job requirements by facility and grade progression.

Tab 3 displays average work week and associated time data for the Medical Corps. The same matrix will be developed by facility to determine if there are differences based on type or size of facility.

This is the first comprehensive occupational analysis conducted of the Navy Medical Department. There are many applications for this data. For example, the Nurse Corps Billet Analysis Project Officer will use the data to (1) compare aggregate respondent data for a position to the actual billet description provided by the facility, (2) determine, by position title, tasks most frequently performed, and (3) recommend minimum education, experience or training subspecialty assigned to specific positions. Secondly, the data analysis will be utilized by the Naval School of Health Sciences (NSHS) during annual curriculum review to determine the need for revision of management courses. The data identifies the needs of the Navy and the courses may be revised based on the information and structures identified through the analysis. Additionally, the data may be used by policy planners to review disparities of task significance within job functions, e.g. the significance of patient care and administrative functions required of primary health care providers. This type of comparative analysis will assist in policy decisions regarding type and allocation of resources required by the Navy Medical Department.

PHYSICIAN TIME DATA MEDICAL NOTAP

	LT	LCDR	CDR	CAPT
AVG WORK WEEK*	69	68	63	61
COMMITTEES	6	7	9	13
COLLATERAL DUTIES	16	23	25	18
WATCHES	97	118	112	89



*Hours per week; others are hours per month

NEW ASCII CODAP TECHNOLOGY: MANPOWER, PERSONNEL, TRAINING APPLICATIONS

Dr. Walter E. Driskill, Chair
Texas MAXIMA Corporation
San Antonio, Texas

The Comprehensive Occupational Analysis Programs (CODAP) were initially developed and implemented by the U.S. Air Force in the 1960's. Since that time the scope of analysis capabilities have been enhanced, based on research results. In addition, CODAP evolved from operating on the 1960 state-of-the-art computer capability to present application on Sperry Univac and, in early 1989, on IBM systems. This panel addresses enhancements occurring in the past two years from operational implementation of new CODAP technology to the experimental development of further analysis methods that soon can be implemented operationally. Also, the last paper addresses the use of CODAP technology and output in the forecasting manpower, personnel, and training (MPT) for emerging weapons systems. The discussant elaborated MPT acquisition needs, specifying a need for development of technology to define knowledge, skill, and ability requirements of tasks. These additional data would be of immense value in forecasting MPT requirements for emerging systems early in the life cycle.

INTRODUCTION TO OPERATIONAL ASCII CODAP: AN OVERVIEW

Johnny J. Weissmuller
Texas MAXIMA Corporation

Joseph S. Tartell
Occupational Analysis Division
USAF Occupational Measurement Center

William J. Phalen
Manpower and Personnel Division
Air Force Human Resources Laboratory

The principal occupational analysis technology in the United States Air Force is the Comprehensive Occupational Data Analysis Programs (CODAP) software system, which has supported a major occupational research program within the Air Force Human Resources Laboratory (AFHRL) since 1962 (Morsh, 1964; Christal, 1974) and an operational occupational analysis capability within Air Training Command's Air Force Occupational Measurement Center (USAFOMC) since 1967 (Driskill, Mitchell, & Tartell, 1980). CODAP technology has spread to U.S. and allied military agencies, and to civilian universities and businesses as well (Christal & Weissmuller, 1988).

THREE PERSPECTIVES OF CODAP

CODAP needs to be understood from several different perspectives. From a psychological measurement point-of-view, CODAP can be conceptualized as a task-based job analysis approach; it is sometimes called the TI/CODAP (Task Inventory/CODAP) approach. TI/CODAP is a set of procedures which focus on the analysis and comparison of individual and group job descriptions and their associated biographical data, as well as on tasks and groups of tasks (task modules) and their associated characteristics.

From an applications perspective, CODAP can also be defined in terms of its significant contributions to the mission accomplishment of the Air Force manpower, personnel, and training (MPT) management functions. These contributions include: providing a data-based approach to evaluating and updating Air Force officer and enlisted classification structures, providing an empirical means of restructuring and redesigning jobs, providing data that have been instrumental in eliminating unnecessary training and in pinpointing specific training requirements, and providing a scientifically sound basis for realigning entry-level aptitude requirements among Air Force career fields (Phalen, Staley, & Weissmuller, 1985).

From a software standpoint, CODAP can be defined as a package of computer programs used to input, process, organize, and report occupational data from inventories of tasks. The CODAP system was developed more than 20 years ago as a software package of some 15 general-purpose occupational analysis programs. However, to keep pace with the rapidly expanding needs of Air Force occupational researchers and analysts, the system was eventually expanded into a somewhat confusing aggregate of more than 60 generic and specialized job analysis programs. Many of the programs were hastily developed in response to various crises, with little time available for producing proper program documentation; some programs

project. Whenever possible, USAFOMC and other users (including R & D contractors) were involved in testing of new system software, as a basis for further suggestions and critiques. In addition, progress reviews and other briefings kept managers informed of the direction and timing of project accomplishments.

In the fall of 1987, orientation classes on ASCII CODAP were provided for USAFOMC programmers and analysts to give them at least a beginning familiarity with the new system (Weissmuller, 1987). Extensive flow diagrams were developed to guide these classes and provide a foundation for further ASCII CODAP training and technical, on-line guidance packages for both analysts and technicians.

In 1988, the new system was implemented at the USAFOMC; after January 1st, every new study was processed under ASCII CODAP. The maintenance of the earlier version of CODAP (known as FIELDATA CODAP) by the AFHRL Information Systems Division was discontinued; a maintenance contract for ASCII CODAP was implemented. An automated trouble report procedure was set up for users to document and communicate system deficiencies. To date, indications are that all problems have been corrected expeditiously and that none of the problems have been serious. Thus, the transition has been relatively free of difficulties and the operational system seems to be performing in accordance with or above expectation.

Even though redesign of the system has now been completed and ASCII CODAP implemented for operational use, advanced development continues - particularly in the areas of new techniques and automated support systems to assist job analysts in developing job and task clusters and to expedite making analysis decisions (Phalen, Weissmuller, & Staley, 1985). In addition, an IBM version of ASCII CODAP is being developed under contract (for 1 March 89 completion). The IBM adaptation is being cooperatively funded by the USAFOMC, the U.S. Army, and the U.S. Navy.

ASCII CODAP SYSTEM

Some of the system parameters of FIELDATA CODAP remain unchanged in the new system. These include the limit of 20,000 cases, the capability to cluster up to 7,000 cases, and the limitations to 26 duties and 66 characters for variable descriptions. Under ASCII CODAP, however, a number of other system limits have been modified. The limit per study on the number of tasks has increased from 1,700 to 7,000; the number of modules from 1,000 to 9,999; the number of tasks per module from 1,726 to 7,000; the number of history variables from 999 to 2,000; and the number of computed variables from 500 to 9,999.

Program names under ASCII CODAP have been improved and standardized in order to enhance analyst-technician communications, as well as to facilitate the identification of programs to use in non-standard analyses. The conventions used in renaming CODAP programs combine two 3-character abbreviations which communicate the actions being taken and the target populations. For example, MEMSEL involves selection of members for groups of interest; GRPJOB involves grouping of cases into jobs; and PRTJOB prints out job descriptions for selected job groups. Some familiar old names have been changed to fit this new philosophy; for example, REXALL is now GRPREL - the reliability of a group (of raters). The few exceptions to the

conventions involve Print Only programs (such as PRTVAR for Printing of Variables) and traditional procedures (such as OVLAP, GROUP, and DIAGRAM). A report listing the new program names and flow diagrams for all programs is now available (Hand et al., 1988).

ASCII CODAP has greater flexibility than FIELDATA CODAP, both in terms of the modularity of its software and in the wealth of options available to the user. The redeveloped software includes the separation of some functions so they can be operated independently. There are, for example, a number of different Print (PRT) programs which are no longer tied to computation programs. This gives greater flexibility in the use of computed files, some of which are used only as input to other programs and need never be printed. ASCII CODAP also has improved interface with other software, both in terms of taking other data files into CODAP and in exporting CODAP data files to other systems. This gives greater flexibility, in that an analyst can make use of many external library programs and is not tied to a single internal system for general mathematical processing. For example, an analyst may make use of several different regression analysis programs on CODAP data files, or can bring into CODAP the results of a regression analysis performed externally. This flexibility also takes advantage of new commercial software developments as they become available, without requiring a revision of CODAP each time.

In terms of using the system, ASCII CODAP permits the continued use of "older" patterns, while also supporting "newer" approaches to the analysis of occupational data. Thus, the more traditional methods of hierarchical grouping and job typing are serviced by the new software, as well as newer approaches which involve non-traditional analysis or automated job or case clustering techniques (JOBTYP and MODTYP). This approach provides great flexibility to users of the system and fully supports analysts throughout a full spectrum of current and future occupational analysis techniques.

ASCII CODAP also has increased efficiency, both in terms of the redeveloped software and in the on-line, self-paced training package on how to operate the system. The software includes improved algorithms for processing which eliminate unnecessary steps in many calculations. Such computational efficiencies make possible a more dynamic use of virtual memory, so that less work has to be done in batch processing mode. By making greater use of direct-access methods, the ASCII CODAP software provides much quicker turn-around for many products and thus facilitates many occupational analysis procedures. Greater efficiency in the computation of job descriptions (1,000 times faster) have made computationally intense programs, such as JOBTYP (automated job typing) and OVLGRP (nonhierarchical clustering), feasible.

The separation of functions and use of more logical naming conventions makes orientation and training on the new system more efficient and effective. Our experience to date has been that programmers and analysts who are familiar with FIELDATA CODAP learn the new system very quickly and adapt to the new terminology fairly easily, particularly as they begin to employ it. In addition, an automatic program "chaining" feature was added so that runstreams of sequenced programs can be generated more efficiently to provide a full set of products needed to begin a study. The new standard flow of initial products is much more efficient in terms of analyst's time and greatly expedites the beginning stages of most analysis

projects. In addition, defaults were modified to reduce or minimize the number and size of products automatically run for initial analyses.

ASCII CODAP involves some new conceptualizations of clustering and the possibilities of multiple clustering approaches. Traditional CODAP clustering involved cases and their similarities; the new approach recognizes the need to cluster tasks as well as cases, or to modify normal case clustering values in terms of other variables. Given the multiple clustering possibilities, the former need for multiple KPATH-ordered data files to represent multiple clustering solutions has been replaced by single "presentation sequence" vectors which are linked to a single case or task data file. We still use the term KPATH to refer to case presentation sequences, but we now have TPATH numbers to refer to task clustering presentation sequences. This distinction helps to communicate the basis (cases versus tasks) for the most common clustering possibilities.

For each cluster solution, ASCII CODAP provides typical cluster output data (presentation sequence, stage numbers, between and within values). The system also provides the ability for modifying (reordering) the presentation sequence (MPATH) based on internal or external factors. Internal reordering of cases or tasks within the confines of hierarchically defined groups might be done on the basis of "number of tasks performed" by each case or "percent members performing" each task, in order to visually enhance the transition points between groups in a PRTVAR (Print Variable) report. A case clustering might be modified externally on the basis of a nonhierarchical reclustering refinement, or a task clustering might be modified to reorder task modules into a more logical order. Each presentation sequence is identified in terms of its own 6-character ID, its set of KPATH or TPATH numbers, and the specifications on which it is based.

The new capability for multiple clusterings creates a need for new terminology to define and specify what is being accomplished. Some of the more important terms are defined as follow:

Module - a set (or group) of tasks.

Cluster Solution - the result of running a hierarchical clustering with a specific set of parameters. This creates a set of stages, each identifying a specific set of objects (cases or tasks) which has specific "Between" and "Within" values.

Presentation Sequence - a "sort" order in which to print objects from a cluster solution; any cluster solution may have many possible presentation sequences.

Module Typing - interpreting a hierarchical clustering of tasks to determine which stages (sets of tasks, or modules) should be used in the final analysis or reporting. Modules also permit summarizing task factor data to a higher order or collecting new data at the module level, to expedite or enhance data interpretation and analysis.

Task Co-Performance - a measure of the degree to which two tasks are performed by the same job incidents (and therefore) should be placed in the same module. Various measures other than the binary "do-don't do" measure exist to define this value.

CONCLUSIONS

The modules approach to occupational analysis embodied in the ASCII CODAP technology represents a significant enhancement of previous analytic capabilities. It is an extremely flexible approach which can be used to mimic existing analysis procedures, which can use new data-based methods for interpretation of task or case clusters, or which can use external data to modify the presentation sequence to improve analysis and reporting. This approach, along with other ASCII CODAP enhancements, form the foundation for continuing experimental work to automate the more routine aspects of occupational data analysis and to provide analysts with decision support aids which can make their work more efficient and timely (JOBTYP for defining and selecting case clusters; MODTYP for defining and selecting task clusters). Programs for analyzing and interpreting case and task clusters will be described in a subsequent presentation in this symposium. Results of the operational implementation of ASCII CODAP at the USAFOMC, as well as the outcomes of ongoing advanced development efforts, indicate that there is a high value in this dual-thrust approach.

REFERENCES

- Christal, R.E. (1974). The United States Air Force occupational research project (AFHRL-TR-73-75, AD-774 574). Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory.
- Christal, R.E., & Weissmuller, J.J. (1988). Job-task inventory analysis. In S. Gael (Ed), Job analysis handbook for business, industry, and government. New York: John Wiley and Sons, Inc. (Chapter 9.3).
- Driskill, W.E., Mitchell, J.L., & Tartell, J.S. (1980). The occupational analysis program - a changing technology. Proceedings of the 22nd annual conference of the Military Testing Association. Toronto, Canada: Canadian Forces Applied Personnel Research Unit.
- Hand, D.K., Haynes, W.R., Staley, M.R., & Weissmuller, J.J. (1988, March). ASCII CODAP: Annotated procedural-level flowcharts for program sequencing. San Antonio, TX: The MAXIMA Corporation (Draft report prepared for AFHRL/MO under AF Contract F33615-83-C-0030).
- Morsh, J.E. (1964). Job analysis in the United States Air Force. Personnel Psychology, 17, 7-17.
- Phalen, W.J., Staley, M.R., & Weissmuller, J.J. (1985, October). Implementation of ASCII CODAP: New versus old system. Proceedings of the 27th annual conference of the Military Testing Association. San Diego, CA: Naval Personnel Research and Development Center.
- Phalen, W.J., Weissmuller, J.J., & Staley, M.R. (1985, May). Advanced CODAP: new analysis capabilities. Proceedings of the Fifth International Occupational Analysts Workshop. Randolph AFB, TX: USAF Occupational Measurement Center.
- Weissmuller, J.J., (1987, October). Introduction to ASCII CODAP for USAFOMC programmers/technicians. Presentation to USAFOMC, Randolph AFB, TX (Training provided under AF Contract F41800-87-N3007).

ASCII CODAP PROGRAMS FOR SELECTING AND INTERPRETING TASK CLUSTERS

William J. Phalen, Air Force Human Resources Laboratory
 Michael R. Staley, ISD Corporation
 Jimmy L. Mitchell, McDonnell Douglas Astronautics Company

INTRODUCTION

Even though revision of the Comprehensive Occupational Data Analysis Programs (CODAP) has been completed and ASCII CODAP implemented for operational use, advanced development continues--particularly in the areas of new techniques and automated support systems to assist job analysts in developing job and task clusters and to expedite making analysis decisions. Within the last two years, considerable progress has been made in developing and validating new technologies for selecting, refining, and interpreting job and task clusters.

This paper will focus entirely on new technologies for defining and interpreting task clusters, because this technology is less well known and understood, even though it has become an increasingly important component of major Air Force Manpower-Personnel-Training (MPT) R&D programs, such as the Training Decisions System (TDS), an Advanced On-the-Job Training System (AOTS), and the Job Performance Measurement (JPM) System, to name but a few. Some basic questions asked by these research efforts have been: What tasks are co-performed? What tasks should be trained together and do co-performance measures accurately and uniformly identify these tasks? How can we meaningfully define a job in terms of a relatively small number of descriptive components, rather than the 500 to 1,500 task statements routinely contained in an occupational survey? Every one of these questions involves the clustering of tasks on an appropriate criterion. A major contribution of the new ASCII CODAP system has been to offer old and new ways of clustering tasks and interpreting task clusters.

PHASE I: THE ADAPTATION OF EXISTING SOFTWARE

Several new approaches to clustering tasks are under consideration and experimentation--most notably, semantic clustering procedures--but the currently preferred approach has been to cluster tasks on a measure called "co-performance." Co-performance can and does mean many things in the ASCII CODAP system. In general terms it refers to measuring the commonality of pairs of task profiles across all the cases in a survey sample. The measure can take on various forms for a pair of tasks, e.g., tasks "A" and "B," which have values across "i" cases:

$$\frac{1}{2} \left[\frac{\sum_i \text{MIN}(A_i, B_i)}{\sum_i A_i} + \frac{\sum_i \text{MIN}(A_i, B_i)}{\sum_i B_i} \right] 100 \quad \left[\frac{2 \sum_i \text{MIN}(A_i, B_i)}{\sum_i A_i + \sum_i B_i} \right] 100$$

$$\left[\frac{\sum_i \text{MIN}(A_i, B_i)}{\sum_i A_i + \sum_i B_i - \sum_i \text{MIN}(A_i, B_i)} \right] 100 \quad \left[\frac{\sum_i \text{MIN}(A_i, B_i)}{\sqrt{(\sum_i A_i)(\sum_i B_i)}} \right] 100$$

The values of "A_i" and "B_i" may be the raw task ratings, or the percent time spent values, or simple, dichotomous "do-don't do" values compared and summed across all cases (N_i) in the sample. When the measure is percent time spent, the A_i and B_i values will be 100% and all four measures shown will reduce to the more simplified expression:

$\{ \text{MIN}(A_i, B_i) \}$. Many other co-performance measures are available in the ASCII CODAP system, but the four listed here suggest some of the possibilities. In actuality, the procedure of choice currently has been to convert the task vectors (rows) to percent time spent values after the case vectors (columns) have been converted to percent time spent values by the INPSTD program and, then, to use the $\{ \text{MIN}(A_i, B_i) \}$ overlap option to compute the co-performance between all possible pairs of tasks. Because the overlap (OVLAP) program works only with columns, this procedure requires front-end application of a program called "XPOSE" to transpose the case-by-task matrix into a task-by-case matrix prior to computing pairwise overlap values. The grouping procedure used for task clustering is a form of average linkage identical to that used for collapsing the overlap matrix when cases are clustered. The resultant hierarchical clustering solution can then be displayed by the DIAGRM program.

Other off-the-shelf software also proved useful for displaying the clustering results. The PRTVAR and PRTFAC programs were used to print the task titles in TPATH order, a product that mimicked the PRTVAR presentation of job titles in KPATH order for cases. The PRTVAR and PRTFAC reports of tasks could also display any other available task variables, such as percent time spent on each task by the total sample, mean learning difficulty and mean recommended training emphasis ratings for each task, etc. PRTVAR was also used to provide a product conveying powerful visual effects; namely, a simultaneous display of a case clustering and a task clustering. As in a standard PRTVAR report, cases would be shown in KPATH order as rows and tasks would be shown in TPATH order as columns (just like background variables). The data items within the cells of this matrix-like report would be the raw task ratings (1-9 relative time spent scale). The visual effect was that of blocks of ratings representing homogeneous clusters of cases (job types) performing homogeneous clusters of tasks (task co-performance modules). When the two types of clustering are displayed simultaneously, each tends to highlight the less obvious breakpoints in the other clustering solution. A planned improvement to enhance the visual effect of the blocks of ratings is to use symbols with varying levels of print intensity to replace the 1-9 ratings as the elements of the matrix.

PHASE II: DEVELOPMENT OF NEW INTERPRETIVE SOFTWARE

To interpret the task clusters derived by analysis of the hierarchical clustering solution, two new programs called "TASSET" (task set) and "CORCAS" (core cases) were developed. These two programs completed a set of four CORSET (sets of core cases or tasks) programs for analyzing and interpreting job and task clusters. Figure 1 shows the interrelationships among the four programs.

Figure 1. The CORSET Programs

	Task Clusters (Task Modules)	Case Clusters (Job Types)
Core Tasks	TASSET	CORTAS
Core Cases	CORCAS	CASSET

Each diagonal has its own naming convention. The TASSET and CASSET programs on the principal diagonal have to do with defining core tasks for task clusters and core cases for case clusters, respectively; while the CORTAS and CORCAS programs on the other diagonal are concerned with defining core tasks for job types and core cases for task modules, respectively. This paper will discuss only the TASSET and CORCAS programs, which are used to interpret task clusters.

THE TASK SET (TASSET) REPORT

Once task clusters have been identified as being distinctly different from one another, the focus moves to pinpointing the nature of the differences. This is the purpose of the TASSET report. It is such a complex report containing so many items of information, we will only be able to touch upon some of the more important items in this paper.

1. Supergroup/Subgroup Matrix. This part of the TASSET report is an asymmetric matrix of percentage values indicating the degree to which each cluster of tasks is co-performed with every other task cluster. Thus, if task cluster "A" is performed, what is the probability that task cluster "B" is performed? And if task cluster "B" is performed, what is the probability that task cluster "A" is performed? From this matrix we can ascertain whether "A" subsumes "B," or "B" subsumes "A," or whether they subsume each other and thus ought to be merged into a single group.

2. Average Co-Performance. We now look at each cluster separately. Within each cluster, TASSET computes the average co-performance of each task with every other task in the cluster. Thus, if a cluster consists of tasks A, B, C, and D, the average co-performance of task "A" would be the average of the co-performances of "A" with "B," "A" with "C," and "A" with "D." This is done for each task in the cluster. The tasks are then sorted high to low on these values. The significance of this measure is that it shows which tasks are most representative of the cluster and which are least representative. Such information is valuable in determining the dominant characteristics of the cluster and in giving the cluster a name.

3. Task Co-Performance Discrimination. TASSET computes a discrimination value for each task in a cluster as a measure of how well each task fits in that cluster compared to how well it might have fit, on the average, if it had been put in each of the other selected task clusters. By this procedure, we can compare task clusters that consist of distinctly different sets of tasks. For example, we have three task clusters: #1 contains tasks A, B, C; #2 contains tasks D, E, F; and #3 contains tasks G, H, I. To compare these three clusters, we would compute the discrimination of task "A" in cluster #1 by evaluating its average co-performance with the tasks in cluster #1 relative to its average co-performance with the tasks in cluster #2 and cluster #3; i.e., assuming task "A" to have been placed in each of the three clusters. If the average co-performance of task "A" is 80% with its own cluster (#1), 40% with cluster #2, and 20% with cluster #3, its discrimination value as a member of cluster #1 would be computed as:

$$DISC_A = \left[\frac{\sqrt{80-40} + \sqrt{80-20}}{2} \right]^2 = 49.5\%$$

A discrimination value of 49.5% indicates that, on the average, task "A" has a 49.5% higher co-performance with cluster #1 than with #2 or #3. Provision is

made for negative discriminations, as well. As a rule of thumb, the cutoff to identify a discriminating task is 25%. Thus, task "A" serves well to discriminate task cluster #1 from task clusters #2 and #3. When TASSET sorts the tasks within a cluster high to low on their discrimination values, a clearer picture emerges of the cluster's unique characteristics. This further helps identify the salient characteristics of the task cluster for describing and naming it.

4. Potential Core Tasks. TASSET lists a set of potential core tasks for each task cluster. Although these are tasks which grouped into other task clusters, they had average co-performance values for the target cluster that were at least as high as one of the tasks actually placed in the target cluster. TASSET also lists the identification numbers (IDs) of the clusters in which the potential core tasks were placed. A large number of potential core tasks with the same cluster ID would indicate a high degree of commonality between this cluster and the target cluster.

5. Uniquely Co-Performed Tasks. Sometimes "uniquely co-performed tasks" are reported by TASSET for a task cluster. These are tasks which are performed by so few people that they do not appear in any task cluster but are reported as unique to a particular cluster if all or most of their co-performance is associated with the tasks in that particular cluster. For example, "Compute complex statistics, such as correlation coefficients and standard deviations," was a task performed by only 1.59% of a sample of clerical workers, but it appeared as a uniquely co-performed task for a cluster of tasks having to do with "compilation of information."

6. Negatively Unique Tasks. Negatively unique tasks are reported by TASSET for a target task cluster when there are tasks which have high co-performance values for most other task clusters (because they are performed by a high percentage of all workers in the occupation) but are rarely co-performed with tasks in the target cluster. For example, clerical tasks related to "dealing with the public" appear as negatively unique tasks for task clusters totally devoted to "transcribing" or "filing."

Various other kinds of task-related information are available in a TASSET report, but the several described here give a good idea of the value of the report for detailed analysis and interpretation of task clusters.

THE CORE CASE REPORT (CORCAS)

In the TASSET report, a task cluster is looked at in terms of its most highly co-performed and most discriminating tasks, so that an analyst can better characterize and name the cluster. Another useful way to characterize a task cluster is in terms of the people who most perform it, and especially those principal performers whose jobs are concentrated in this task cluster to the exclusion of all or most other task clusters. This is precisely what the CORCAS report does. The CORCAS report may contain any type of background variable information describing the case that will fit in the allocated space, just as on a PRIVAR report; however, "job title" is often the most useful variable. By way of example, suppose an analyst was trying to understand why a task cluster consisted of a set of nondescript clerical tasks. By applying the CORCAS program to the task cluster, he found that most of the cases who are most heavily involved with this task cluster listed as their job title "library clerk." This would indicate that this set of tasks (which nowhere mentions the word "library"), is peculiar to clerical work in a library setting. Other variables, such as grade level, organization type or

level, official job codes, and even KPATH sequence numbers (job type relationships) can be useful interpretive indicators in coming to an understanding of "why" a task cluster occurred in terms of "who" performs it.

THE JOB TYPE VS TASK MODULE MAPPING REPORT (JOBMOD). The JOBMOD program aggregates the case- and task-level indices computed by the four "CORSET" programs and uses these aggregate measures to relate task clusters to job types and vice versa. The description of job types by a handful of discriminant clusters of tasks and the association of each task cluster to the types of jobs, of which it is an important component, is a basic requirement for defining and integrating the MPT components of an existing or potential Air Force Specialty (AFS) or weapon system. If AFSs are to be collapsed or shredded out, or new jobs are to be assigned to an occupational area, or old jobs are to be moved to another occupational area, such highly summarized, yet meaningfully discriminant hard data are essential. There is an especially critical need for this kind of data if numerous AFSs are to be reviewed and compared by panels of subject-matter experts (SMEs) or functional managers who must make broad-based judgments, recommendations, or decisions. JOBMOD products will provide a welcome assist to those who are responsible for meeting another important MPT requirement, that of integrating task data from a variety of databases, such as the Logistics Composite (LCOM), the Logistics Support Analysis (LSA), Maintenance Data Collection (MDC) and the Occupational Survey (OSk) databases. Task clusters and their associated job types provide fairly stable, functionally discrete categories into which tasks from other data sources can be mapped and ultimately integrated.

PHASE III: DEVELOPMENT OF AUTOMATED CLUSTER SELECTION AND REFINEMENT SOFTWARE

The MODIYP Program

Just as the JOBTYP program automatically selects the "best" set of job types from a hierarchical clustering of cases, based on similarity of time spent across tasks, the MODIYP program selects the "best" set of task module types from a hierarchical clustering of tasks, based on task co-performance across cases. The term "best" means that the evaluation algorithm initially optimizes on four criteria simultaneously, i.e., within-group homogeneity, between-group discrimination, group size, and drop in "between overlap" in consecutive stages of the hierarchical clustering. After all stages of the clustering have been evaluated on these criteria, a primary, a secondary, and a tertiary set of mutually exclusive task clusters are selected as first-, second-, and third-best representations of the modular structure of the hierarchical clustering solution. The three sets of groups are then input to another evaluation algorithm which computes supergroup and subgroup indices between all pairs of groups in the primary solution, between secondary groups and primary groups within the same KPATH range, and between tertiary and secondary groups, as well as between tertiary and primary groups, within the same KPATH range. Based on the combined results of both evaluations, the primary, secondary, and tertiary sets of groups are revised, i.e., groups at one level may be promoted or demoted to replace a group or groups within the same KPATH range at another level. The final set of primary groups is input to the TASSET and CORCAS programs to provide analytic and interpretive data for each primary cluster of tasks. The MODIYP output also produces a report which shows the initial and final sets of primary, secondary, and tertiary groups and their evaluation indices. In the several applications of MODIYP to date, MODIYP has picked virtually the same task clusters as the analysts in one-third to one-half of the selections, and has deviated by no more than one or two clustering stages in another one-third to one-half of the selections.

In less than one-sixth of the selections has MODTYP deviated radically from the analyst selections and, in those instances, it was usually arguable as to which was the better selection. When the analyst deviates significantly from the MODTYP selection, it is hypothesized that the upward or downward location of the secondary and tertiary groups on the DIAGRM report will be predictive of the direction of deviation. Insufficient data have been analyzed at this point to warrant acceptance or rejection of this hypothesis. The special value of automated cluster selection programs like JOBTYP and MODTYP in an MPT research environment is that they provide completely automated, rapid, yet reasonably accurate and highly standardized results that are completely replicable. Accuracy, standardization, and replicability of results across multiple occupational survey studies is essential if MPT integration is to be a systematic process. Where occupational analyses have been done by inexperienced analysts, or where the principal objective of the analysis was to identify career field structure at the major cluster level, rather than at the job type level, the completely automated feature of these programs becomes critical, because the MPT researcher may not be equipped to perform a standard CODAP job type analysis. The rapidity of the procedure will play an important role if a number of job or task module typing exercises must be done in order to proceed to the next step in the MPT integration process.

The JOBMOD program and its role in MPT integration was described earlier in this paper. If JOBTYP and MODTYP are employed to produce the job and task clusters to be input to JOBMOD, then JOBMOD becomes the end product of a totally automated process that may well serve as the appropriate beginning analysis/evaluation point for the MPT integration process.

THE OVLGRP AND SEDGRP PROGRAMS

If JOBMOD provides the starting point for the MPT integration process, one might ask whether it could be made a better one. Could the job and task clusters consist of a better balance of within-group homogeneity, between-group difference, and number of cases or tasks classified into the selected clusters? The answer is affirmative, and the refinement procedure to be used to accomplish this balancing act is the OVLGRP nonhierarchical clustering procedure. OVLGRP can use the centroids of hierarchically formed clusters of jobs or tasks as input seeds around which to optimally cluster all cases or tasks, or it can use the input seeds generated by the SEDGRP program, i.e., discriminant cases or tasks which represent the entire measurement space encompassed by the sample data. Specific applications of the nonhierarchical clustering procedures will be discussed at length in another paper presented at this symposium. OVLGRP and SEDGRP are mentioned here in order to indicate how job and task clusters can be refined prior to their being input to JOBMOD.

CONCLUDING REMARKS

The purpose of this paper was to familiarize actual and potential ASCII CODAP users with a number of ASCII CODAP products specifically designed for the definition, selection, analysis, and interpretation of task clusters. These products allow clusters of tasks to be examined, compared, and interpreted with the same degree of care and meticulousness previously associated only with case clustering and the resultant job clusters. The USAF Occupational Measurement Center (USAFOMC) is already using the new ASCII CODAP technology to combine and integrate the two analytical streams in an overall multidimensional approach to studying the world of work. Future technology R&D efforts will attempt to improve and more fully automate this integration process in support of the MPT integration research effort.

OPERATIONAL TESTING OF ASCII CODAP
JOB AND TASK CLUSTERING METHODOLOGIES

Jimmy L. Mitchell
Systems Engineering & Analytics Department
McDonnell Douglas Astronautics Company

William J. Phalen
Manpower and Personnel Division
Air Force Human Resources Laboratory

William Haynes
Metrica, Inc.

Darryl Hand
Texas MAXIMA Corporation

INTRODUCTION

The ASCII CODAP refinements developed to enhance occupational analysis capability have been operationally tested on a number of example data sets representing several recently completed occupational analysis projects. Using such examples, the operational testing compared several algorithmic solutions with those actually made by experienced analysts. Feedback from such tests was used to further refine and adjust the algorithms used to identify potential job and task clusters. New displays and adapted CODAP products needed for an analysis to make final job type or task module decisions were developed and their utility tested in actual use.

REFINEMENT OF JOB AND TASK GROUPINGS

In a typical occupational study, groups identified as meaningful jobs that are interpreted from a diagram of an occupation will not include all the cases in the sample, except at a very low stage, where the overlap values are low. In previous reports (Mitchell and Phalen, 1985; Phalen, Mitchell, and Staley, 1987), we reported the use of an iterative nonhierarchical cleanup procedure to solve this problem and refine the groups. A sample of the results can be seen in Figure 1, where groups from the hierarchical clustering account for 80% of the cases and the refined groups output by the nonhierarchical refinement process encompass 99.2% of the cases (in iteration 6). By computing the percent time overlap of each case's job description with the mean description for every selected group, this procedure permits each unclassified case to be included in the group it most resembles, at a specified minimum level of overlap. The group vector is then recomputed and, in the second iteration and beyond, all cases are compared to the group means. Cases can migrate to the emerging

ITERATION	NO. CLASSIFIED	NO. UNCLASSIFIED	PERCENT
Diagram	2643	660	80.0
1	3298	5	98.8
2	3292	11	98.6
3	3285	18	99.4
4	3281	22	99.3
5	3279	24	99.3
6	3278	25	99.2

Figure 1. OVLGRP - Totals by Iteration (AFS 811XX; data from Alton, 1984).

groups they are most like, rather than being forced to remain with the first linkage, as is the case in hierarchical clustering.

We need to be able to trace some example groups through some of these iterations in order to see what is happening to them. Figure 2 shows three Law Enforcement groups to illustrate the type of changes which occur. In this case, we can track where the people go since the 22 members lost from the combined Desk Sgt/Patrol group are the same 22 showing up in the Desk Sgt group. Note that once these cases are added, that group tends to stabilize in terms of size and the variance in overlap of individuals with the group mean drops considerably (S.D. = 9.2 --> 7.3), which is evidence of a more homogeneous group. The LE Patrolmen group increases in size and decreases in variance, while the third group does the opposite. We end up with two meaningful groups instead of three, and their job descriptions should be more realistic pictures of their jobs.

GROUP	VARIABLE	DGRM GP	ITERATION					
			1	2	3	4	5	6
GRP 594	LE DESK SERGEANT							
	Grp Size	15	37	34	33	32	32	33
	No. Lost	-	-	3	2	1	0	0
	Gained	-	22	0	1	0	0	1
	Mean Ovrtp	-	44.6	42.5	42.7	42.4	42.5	42.6
	S. D.	-	9.2	7.4	7.6	7.4	7.3	7.3
GRP 921	LE DSK SGT/PATROL							
	Grp Size	90	68	38	20	14	10	5
	No. Lost	-	22	30	18	7	4	5
	Gained	-	0	0	0	1	0	0
	Mean Ovrtp	-	57.4	56.2	56.9	55.3	54.4	49.2
	S. D.	-	7.2	7.4	8.2	9.3	10.5	10.9
GRP 785	LE PATROLMEN							
	Grp Size	12	38	55	90	84	67	51
	No. Lost	-	-	6	12	29	25	20
	Gained	-	26	23	47	23	8	4
	Mean Ovrtp	-	51.0	52.1	55.5	57.2	57.6	57.9
	S. D.	-	10.9	7.8	7.4	7.2	6.9	5.9

Figure 2. OVLGRP - Details of Example Law Enforcement Groups.

The reclassification of cases in iterative stages is done with OVLGRP (i.e., the overlap of cases with group means), which identifies new groups of cases which are more internally consistent and may be somewhat easier to interpret in terms of core or characteristic and distinguishing tasks. The interpretation of the groups at each iteration is not an easy task but better reports are now available to track how cases move among job groups and to analyze new job groups.

A GRPMAT (a table which shows the migration of cases from one job group to another) for the Precision Measuring Equipment Laboratory (PMEL) study is shown as Figure 3. Not all of the 21 OSR groups (Aslett, 1984) are shown in this display in order to simplify the discussion. Note that the job types are shown down the lefthand column and the OVLGRP-refined groups across the top of this display. Those cases not members of any group in OVLGRP are shown as the first column (GRP 02), and the marginal summaries (ST 001) report the size of each group. These data provide some examples of what happens in the reclustering process.

-----These are the Job Types from the OSR

These are the Job Types after OVLGRP								
v	: Grp 02	Grp 03	Grp 04	Grp 05	Grp 06...	Grp 09	Grp 10...	ST001
ST 141	.	95	2	.	1	.	6	112
ST 096	.	2	160	.	7	7	7	211
ST 322	.	.	.	49	6	11	.	181
ST 238	17	1	.	19
ST 140	82	4	90
ST 243	2	3	5
.								.
.								.
.								.
ST 001	63	103	194	49	39 ...	120	62 ...	1513

Figure 3. GRPMAT - Precision Measuring Equipment Laboratory (FMEL) Personnel.

Note that on the diagonal most of the major groups have retained their identity. Of 112 cases in Stage 141, 95 remain along with eight picked up from other groups; this represents relatively little change and thus there will be no real change in the original job description. For Stage 96, 160 of the original 211 remain, and an additional 34 are added from other groups. A core of 49 people from the original 181 in Stage 322 stayed together and added no new members; the remaining 132 cases were scattered across eight other groups.

Stage 238 was a small group of 19; core tasks which discriminate this group have to do with calibrating HF counters, align or troubleshoot electronic counters or spectrum analyzer plug-in units. Seventeen of the 19 cases provided a stable core around which another 22 cases clustered. Examination of the tasks for the new group indicates that the tasks which discriminate the group remain the same, but the percent of the group performing the core tasks has increased.

In addition to seeing the numbers of individuals who move or stay in the various groups, an analyst also needs to examine the background data for these individuals to fully understand the significance of the various clusterings. Summaries of such data in clustering sequence (PRIVAR) are routinely used in the normal analysis process. Since members have migrated among groups, the normal diagram sequence is no longer valid; one would have to trace individuals through two sequence numbers case by case. To meet this need, a new PRIVAR option is available which provides a separate product for each of the new job groups and which displays the data in original clustering (KPATH) sequence.

Figure 4 reproduces a portion of a PRIVAR for GRP 003 (original Stage 96); the original clustering sequence for the group was 113 to 323. Much of the original KPATH sequence remains intact (in fact, 160 of 211 cases), with 2 cases added from below the original range. In the upper KPATH sequence, there is more mixing with some cases migrating in or out. It remains predominately a first job with about half the members holding a 3-skill level AFSC. There are, however, some 7-skill level personnel in the group; they perform about the same number of tasks and have a technician title as opposed to calling themselves supervisors or NCOICs.

KPATH	NO.TASKS	AFSC	GRADE	MAJCOM	BASE	JOB TITLE
100	95	32470	TSgt	SYS	Hanscom	FMEL Technician
102	76	32430	A1C	USAFE	Torrejon	FMEL Technician
113	73	32450	A1C	SAC	Andersen	Prec Meas Equip
114	88	32430	A1C	USAFE	Ramstein	PME Spec
115	106	32430	A1C	ATC	Williams	FMES
116	72	32430	A1C	MAC	Little Rk	FMEL Technician
117	79	32450	Sgt	SAC	Andersen	Precision Meas.
118	80	32430	A1C	SAC	Grnd Fks	FMEL Specialist
119	64	32430	A1C	ATC	Lowry	PME Spec
.
.
.
319	74	32470	TSgt	SAC	Wurtsmith	K 1-9 Sec Spvr
321	108	32470	SSgt	USAFE	Lakenhth	PME Tech
324	77	32450	SSgt	MAC	Bolling	PME Tech
325	63	32470	SSgt	OAR	HQ AF cen	FMEL Mobile Cal
.
.
.
1093	47	32430	A1C	AAC	Shemya	FMEL Tech

Figure 4. PRTVAR - GRPO03; Precision Measuring Equipment Personnel.

Missing cases (that is, those which migrated out from the group) differed from rest of the group in terms of their relative time spent on duties. They tended to migrate out in pairs which had some similarity of duty time with the group but usually performed more duties as well.

We need not get totally submerged in the details of this process here. The point is that the reclustering of groups does help to identify distinguishing tasks of various jobs and can be used by an analyst to refine his or her initial job type selections. Major groups are relatively stable in terms of their job descriptions, but some of the smaller groups proved unstable and disappear. Other small groups involved in specialized missions, such as F-15 equipment maintenance, not only proved stable but increased in size in this iterative regrouping process.

The intense analysis work involved in interpreting and evaluating the regrouped job types can be facilitated by using CORIAS and PRTVAR outputs for the groups which need to be reanalyzed (those which were not stable as indicated in GRPMAT, or where the analyst wishes an improved job description). This approach reduces the total amount of effort needed to reanalyze the set of AFS jobs. Another approach uses AUTOJET runs to compare pairs of groups an analyst wishes to study; such AUTOJET products help to expedite comparing two groups (input versus output groups) and provides a very quick way to highlight the differences between such groups.

OPERATIONAL TESTING OF THE NEW APPROACH

To test out the new procedures, ten third-level personnel in 10 job and task clusters, six Air Force specialties, were examined using this approach. Results were quite good and, in general, reflected the "normal" analysts' judgments very closely (some slightly more specific, and slightly more general).

The very positive results from the validation testing led us, in one of the monthly CODAP Users' meetings, to volunteer to use the procedure for any study where the USAFOMC analyst was having difficulty with an analysis or where more than 10% of the cases were not covered by the identified job types. USAFOMC personnel suggested that a good candidate specialty would be the Supply AFS, where almost 50% of the cases were not included in the analyst's final job types (Caussade 1988).

Initial trials using the Supply AFS jumped the coverage of cases to 96.9%, which seems an excellent outcome (see Figure 5). However, close examination of the GRPMAT, PRIVAR, and OVLGRP products showed an increase in standard deviations and decrease in mean within group overlap across iterations--exactly opposite of expectation.

ITERATION	NO. CLASSIFIED	NO. UNCLASSIFIED	PERCENT
Diagram	1958	1793	52.2
1	3604	145	96.1
2	3624	127	96.6
3	3627	124	96.7
4	3631	120	96.8
5	3634	117	96.9
6	3632	119	96.8

Figure 5. OVLGRP - Total in Each Iteration; July 1988 (AFS 645XX).

Further analysis of the Supply (AFS 645XX) products suggested that while the unclustered cases were now included in identified groups, the groups were mostly at an unacceptable level of internal overlap to be considered valid job types. Clearly, the program was not operating appropriately for us achieve the end for which it was designed. This unexpected outcome led to a complete reexamination of the software.

The problem appears to have been with routines which set the minimum acceptable overlap before a case could be considered a member of a group; the minimum cutoff was not operating as desired. Thus, all cases were being merged with some group, the one which was the "best" fit for the case when compared with all other groups. This routine was rewritten to effect a dual minimum overlap cutoff criteria, one that is absolute and one that varies from group to group. If a case does not meet the dual criteria, it is rejected for membership in any group (and will be held for reconsideration at the next iteration).

Figure 6 displays the results of the revised program. In this case, the coverage of cases has been improved from about 52% to about 71% of the cases, without degrading the job descriptions of the groups. The remaining 29% of the cases were still so divergent that they are not included.

The resulting groups are much more acceptable and have some face validity. One result was the breakup of one large group of supervisors into smaller groups more directly related to the various technical areas within the specialty. The USAFOMC analyst involved in this study indicated that this made much more sense in that the smaller groups could be directly related to the technical subareas.

Certainly, the 71% coverage is less than desired, but it may be all that is possible with automated technology for heterogeneous specialties. When viewed in terms of improvement over the original clustering, it represents a significant gain in group coverage (19%). In addition, the final job groups appear realistic and more interpretable to the analyst.

ITERATION	NO. CLASSIFIED	NO. UNCLASSIFIED	PERCENT
Diagram	1958	1793	52.2
1	2509	1242	66.8
2	2571	1180	68.5
3	2648	1103	70.6
4	2664	1087	71.0
5	2669	1082	71.2
6	2677	1074	71.4

Figure 6. OVLGRP - Total in Each Iteration with Improved Program;
October 1988 (AFS 645XX).

For most studies, the improved program should provide for 95 to 99% coverage of cases. There will be some studies (such as Supply), however, where the AFS is so diverse (or where the jobs are not well structured) that even with the improved program, coverage can not be more than 70 - 75%. We would maintain that such complex or heterogeneous AFSs are in need of close examination. They may represent areas for possible reengineering efforts (shredouts or functional reorganization).

CONCLUSIONS

Operational testing of the nonhierarchical refinement of job types has demonstrated that in most AFSs, very significant improvements can be made in the numbers of members covered by identified job types, and calculated job descriptions will be more stable. For some very diverse AFSs, this refinement procedure will help, but will not completely solve the problems which such diversity of jobs represent. This refinement technique is another very valuable tool which should help analysts improve and expedite their work.

References

- Alton, R. L. (1984, November). Occupational survey report, Security Police career ladders (AFS 811X0, 811X2, and 811X2A). Randolph AFB, TX: USAF Occupational Measurement Center.
- Aslett, L. S. (1984, February). Occupational survey report, Percision Measuring Equipment career ladder (AFS 324X0). Randolph AFB, TX: USAF Occupational Measurement Center.
- Caussade, Jose. (1988, July). Occupational survey report, Supply career ladder (AFS 645XX). Randolph AFB, TX: USAF Occupational Measurement Center.
- Mitchell, J.I. & Phalen, W.J. (1985, October). Non-hierarchical clustering of Air Force jobs and tasks. Proceedings of the 27th Annual Conference of the Military Testing Association. San Diego, CA: Naval Personnel Research and Development Center.
- Phalen, W.J., Mitchell, J.L., & Staley, M. R. (1987, May). Operational testing of ASCII CODAP job and task clustering refinement methodologies. Proceedings of the Sixth International Occupational Analysts Workshop. San Antonio, TX: USAF Occupational Measurement Center.

ASCII CODAP AND MANPOWER-PERSONNEL-TRAINING (MPT) TECHNOLOGIES

R. Bruce Gould and Hendrick W. Ruck, Air Force Human Resources Laboratory
Walter E. Driskill, Texas MAXIMA Corporation
Jay S. Tartell, USAF Occupational Measurement Center

INTRODUCTION

The initial building block for most Manpower-Personnel-Training (MPT) analysis technologies is task-level job information. Of the variety of task measures available, four are primary for use in MPT technologies: tasks performed, time spent performing specific tasks, training emphasis (TE), and learning difficulty (LD). Over the past 25 years, the Air Force has evolved an approach for reviewing personnel classification and utilization policies and for determining technical training content based on tasks performed (Morsh, 1964; Christal, 1974; Mitchell, 1988). As part of the normal occupational analysis (OA) process, tasks are defined by subject-matter experts (SMEs) of a specialty in their own technical terminology, working with analysts of the USAF Occupational Measurement Center (USAFOMC) (see AFR 35-2).

Several kinds of data on these tasks are collected from job incumbents and supervisors for use in reviewing training programs. Large samples of incumbents are asked to provide information about which tasks they perform in their present jobs and the relative amount of their job time spent performing such tasks. These data are used to examine the variety of specialized jobs within an Air Force Specialty (AFS), to assess how jobs change at advanced skill levels, and to review official specialty descriptions and initial training programs (Christal & Weissmuller, 1988; Mitchell, Ruck, & Driskill, 1988).

TE ratings have been validated empirically using explanatory regression models in studies of 18 AFSs (Ruck, Thompson & Stacy, 1987; Ruck, Thompson, & Thomson, 1978). Two important findings of these research studies were that supervisors agreed substantially with one another on their recommendations in most career fields, and that the supervisors' judgments were explainable in terms of key Instructional Systems Development factors. A third important finding was that the supervisors could not agree as to the appropriate sites for training technical tasks. TE ratings are used primarily to evaluate course content of basic technical training courses (Mitchell et al., 1988).

The learning difficulty (LD) technology has provided a reliable means for establishing the relative learning requirements of tasks within specialties and, when used with a benchmarking process, the relative LD across specialties. These relationships are currently used to set specialty minimum aptitude standards and in the maximizing function for classifying airmen into jobs through the person-job-match system. The LD benchmarking process requires that experts observe samples of tasks being performed, so it is very time and resource intensive. A more economical procedure is under development and is reported in another conference paper (see Davis, 1988). Whereas the current occupational learning difficulty (OLD) process requires observing task performance, the new procedure requires only SME ratings of task descriptions. The implications are that the process can now be used to establish learning and aptitude requirements for operator and maintenance jobs associated with weapon systems which are still on the designers' drawing board rather than just used with existing, operating systems.

Access to the measures described above is provided in part by the Occupational Research Data Bank (ORDB). The ORDB is a user-friendly, on-line system which combines CODAP data for all major AFSSs with specialty requirements documents and reports, and 125 demographic variables from personnel files. ORDB provides ready access to MPT-relevant information by managers and planners for existing weapon systems. Potentially, developers of new weapon systems could obtain MPT-relevant information on both predecessor and comparable systems. Another conference paper outlines the use of the ORDB with MPT technology (Longwire & Short, 1988).

PRESENT STATUS OF ASCII CODAP

The CODAP analysis capability which makes the OA data usable by MPT technology has recently undergone redevelopment. An improved CODAP system was started in 1984, completed in 1987, and is considered to be an extremely successful project. Earlier presentations in this symposium have highlighted some of the new developments of the system and demonstrated some of its potential for further efficiencies and saving of analysts' time. OA data are now used in Utilization and Training Workshops for major training and personnel decisions such as Modification of Specialty Descriptions (see AFR 39-1) and Specialty Training Standards (AFR 8-13). USAFOMC analysts were trained on ASCII CODAP in late 1987, even as the final programs were being developed. The speed with which ASCII CODAP has been accepted and used, both in other R&D programs and in the operational OA program is rather remarkable.

Presently, ASCII CODAP is fully operational only on UNISYS (Spercy-UNIVAC) computer equipment, including the system at AFHRL, Brooks AFB, Texas and one at the U.S. Army Soldier's Support Center in Alexandria, Virginia. As a result of interactions and communications among potential users, another project is also now underway to adapt ASCII CODAP for use on IBM equipment. This work is being done by the MAXIMA Corporation and is funded jointly by the USAFOMC, the U.S. Army, and the U.S. Navy.

The concept of operations in this adaptation of ASCII CODAP is to ensure that all functions are paralleled in all three versions of the software. This will permit any changes (innovations or improvements) to be easily adapted and available in the other operating systems very quickly. In this way, the system can be operationally maintained and updated in all locations with the latest version of the software immediately after some innovation has been developed and validated. This approach will ensure that various users are never limited to some older or obsolete version of the system, as has been the case in the past. The project is scheduled for completion early in 1989, and should result in the rapid spread of the use of ASCII CODAP throughout the broader occupational analysis community.

THE ROLE OF ASCII CODAP IN MPT RESEARCH

ASCII CODAP has a number of enhanced capabilities which have already been used to support several advanced MPT research efforts. In one project, the new system was integrated into the EAF project to take advantage of the advantage of such capabilities.

The Training Research and Development Center (TRDC) is currently conducting training decision analysis research. The TRDC is currently conducting research

appropriate training settings for specialty tasks (task modules) and assess the probable consequences of proposed changes to the career field in terms of training costs and capacities (see Vaughan et al. 1988). ASCII CODAP was selected for use in the TDS to ensure future compatibility with the operational OA program. ASCII CODAP software was used for clustering existing job tasks into efficient task modules for use in the TDS, and the MODULES software was modified slightly to accommodate TDS needs. This application of CODAP was very successful and resulted in the slight modification of existing programs, as opposed to writing entirely new software for the initial TDS subsystem (Perrin et al. 1987; Vaughan et al. 1988).

A second task-based project was the Advanced On-the-Job Training System (AOTS), which provided for the use of generic positions as a basis for defining and managing OJT. Traditional CODAP analysis reports were used as a starting point for identifying the jobs and the development of generic and position-specific Master Task Lists. CODAP data for each target job in the AOTS specialties were used with other AFHRL-developed algorithms to compute first-term training priority scores to help rank training requirements (Datko, Cassidy, & Ruck 1982). The AOTS is a training management and delivery system aimed at helping functional managers and supervisors identify appropriate OJT requirements and assure that such training is delivered to the airmen who need it (Blackhurst & Sturdevant, 1987).

CODAP data for selected specialties also served as the starting point for development of job performance measurement (JPM) instruments to obtain criteria for enlistment aptitude requirements validation or training evaluation. This very successful project permits the examination of the relationship between Armed Forces Vocational Aptitude Battery test scores and job performance. A model of the performance measurement process was developed to guide the JPM research (Kavanaugh, Borman, Hedge, & Gould, 1987). A separate model of job performance highlighted the need for a common basis for job performance assessment and training evaluation (Driskill, Mitchell, & Ballentine, 1985).

CODAP data through the ORDB were used to reorganize specialties in the RIVET Workforce effort. Here the Air Force discovered that deploying aircraft to dispersed operating locations required too many support personnel to make this operational requirement supportable. Further, new weapon systems were requiring more support personnel than the systems they were replacing. Finally, strength reductions meant there were not going to be enough personnel to support existing systems. The solution was to move away from the specialist and toward a generalist structure for specialties. To that end, RIVET Workforce used the CODAP data to structure the new generalists specialties. RIVET Workforce basically identified and collapsed specialties with significant communality as a short-term fix. A new research effort is underway to start with task-level data and cluster tasks into efficient jobs, and jobs into efficient specialties. New ASCII CODAP programs such as non-hierarchical clustering and automated job typing programs will be core to the process of restructuring existing weapon system specialties. The research is building a Specialty Structuring System (S³) which will also be used to construct support specialties for new WSS using engineering task-like data long before those WSS are depleted.

S³ is also important because it addresses one of our greatest problems in using CODAP data to support MPT: integrating M, P, and T decision processes. Typically, the manpower, personnel, and training communities make decisions which affect each other but either don't coordinate those decisions or don't know how to assess the impacts across concerns. For example, numbers of maintenance personnel, required aptitude level, and training time are highly interactive. Building a highly complex job may permit using a smaller staff, but that staff must be smarter and better trained to perform adequately. S³ will consider the tradeoffs between the M, P, and T requirements when structuring jobs. This involves projecting the numbers and kinds of people and training resources required to support emerging WSs, assessing the ability to field and maintain WSs once deployed, and then integrating requirements seeking a supportable life-cycle cost.

Experimental semantic analysis techniques show great promise in merging ASCII CODAP-developed data bases with other source task data such as Logistics Support Analysis Records, thus demonstrating the potential for much broader applications of ASCII CODAP methodologies. This work is being done under the Task Identification and Evaluation System (TIES), which attempts to optimize the use of existing task data bases through proper interaction of automated systems. Results to date are extremely promising and a number of new techniques have been developed to assist in achieving this objective. This data may, for example, be used in early determination of training, manpower, aptitude, and recruiting requirements during the early developmental stages of an emerging weapon system. Consider how such information might positively affect design decisions and limit the cases where the services learn WSs are not supportable after deployment.

THE FUTURE

One primary area where further ASCII CODAP development has great potential for assisting Air Force MPI managers is in the area of new weapon system development and the specification of the MPI requirements for such new systems. Present ASCII CODAP is designed to service the need for OA of current AFSs and does not project what such AFSs "should be" or "could be" in the future. The TDS R&D has demonstrated the capability to model projected changes in AFS jobs and training programs and evaluate the resulting training costs and capacities, but this capability is restricted by the data base to present AFS tasks. Likewise, the S³ uses existing data bases to project AFS requirements.

What is needed is some new technology for predicting future tasks and systematically building estimates of expected MPI requirements (including OJT requirements via the AOTS). We also need to be able to model what effect changes in the manpower pool (aptitude levels, technological background requirements, length of enlistments, and other factors) will have on MPI requirements. Major progress toward such goals will probably involve some merging of TDS, AOTS, S³, TIES, and ASCII CODAP capabilities into an expanded MPI requirements specification and management system.

We also need better methods for flowing task data from one system to other systems which can or should make use of such data. Task data developed during new weapon system development should flow into the Air Force training and personnel systems in automated form where it should be matched to existing AFS

data and evaluated in terms of potential MPT changes required. Experimental semantic analysis techniques should help in the matching process, but further R&D is needed for techniques to evaluate and integrate such data. Linkages must also be built to interface such data with the manpower system to project needed adjustment in authorizations and assess manpower constraints on personnel and training. Likewise, results from such MPT projections need to be linked back to the task list development processes of the USAFOMC for use in formalizing new specifications (AFR 39-1) and updating AFS training requirements.

A critical issue in the evolution of future MPT systems is to be sure that all R&D personnel, as well as MPT staff agencies (AIC, USAFOMC, and AFMPC), functional managers, and other involved personnel, are all highly aware of our common future directions and goals. As the necessary R&D projects develop new MPT technologies, we must be able to ensure they are mutually compatible and plan for their eventual integration (Ruck & Mitchell, 1987). A truly integrated MPT system is not possible without such general awareness and preplanning. ASCII CODAP, with its new flexibility and adaptability, can play a key role in attaining such goals, as fielddata CODAP has done in the past. It provides the needed foundation (task-based data) which will drive such eventual MPT systems integration.

REFERENCES

- Air Force Regulation 8-13 (1980, 12 June). Air Force training standards. Washington, DC: Headquarters, United States Air Force.
- Air Force Regulation 35-2 (1982, 23 July). Occupational analysis. Washington, DC: Headquarters, United States Air Force.
- Air Force Regulation 39-1 (1981, 30 April). Airman classification. Washington, DC: Headquarters, United States Air Force.
- Blackhurst, Jack, and Sterdevant, W.S. (1987, April). Development of an advanced on-the-job training system. Paper presented at the Second Annual Midyear Conference of the Society of Industrial and Organizational Psychology, Atlanta, GA.
- Christal, R.E. (1974). The United States Air Force occupational research project (AFHRL-TP-73-75, AD-774 574). Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory.
- Christal, R.E., & Weissmuller, J.J. (1978). Job-task inventory analysis. In S. Gael (ED), Job analysis handbook for business, industry, and government. New York: John Wiley and Sons, Inc. (Chapter 9.3).
- Datko, L.M., Cassidy, M.J., & Ruck, H.W. (1982, September). Standardized Position Oriented Training System (SPOTS); Task listing generation procedures (AFHRL-TP-82-16, AD-A120 119). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

- Davis, P. A. (1988). A practical procedure for determining occupational learning difficulty. Paper presented at the 30th Annual Conference of the Military Testing Association. Washington, DC.
- Kavanagh, M.J., Borman, W.C., Hedge, J.W., & Gould, R.E. (1987, September). Job performance measurement in the military: A classification scheme, literature review, and directions for research (AFHRL-TR-87-15). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Longmire, K. M., & Short, L.C. (1988). The occupational research data bank: A key to MPT support. Paper presented at the 30th Annual Conference of the Military Testing Association, Washington, DC.
- Mitchell, J.L. (1988). History of job analysis in military organizations. In S. Gael (Ed), Job analysis handbook for business, industry, and government. New York: John Wiley and Sons, Inc. (Chapter 1.3).
- Mitchell, J.L., Ruck, H.W., & Driskill, W.E. (1988). Task-based training program development. In S. Gael (Ed), Job analysis handbook for business, industry, and government. New York: John Wiley and Sons, Inc. (Chapter 3.2)
- Morsh, J.E. (1964). Job analysis in the United States Air Force. Personnel Psychology, 17, 7-17.
- Perrin, B.M., Vaughan, D.S., Yadrick, R.M., Mitchell, J.L., & Knight, J.R. (1986, 7 February). Development of task clustering procedures (Technical Report, CDRL 7B). Brooks AFB, TX: Prepared for the Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Ruck, H.W., & Mitchell, J.L. (1987, April). Integration of training R&D into the USAF MPI system. Paper presented at the Second Annual Midyear Conference of the Society of Industrial and Organizational Psychology, Atlanta, GA.
- Ruck, H.W., Thompson, N.A., & Stacy, W.J. (1987). Task training emphasis for determining training priority (AFHRL-TP-86-65). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Ruck, H.W., Thompson, N.A., & Thomson, D.C. (1978, October-November). The collection and prediction of training emphasis ratings for curriculum development. Proceedings of the 20th Annual Conference of the Military Testing Association. Oklahoma City, OK: U.S. Coast Guard Institute.
- Vaughan, D.S., Mitchell, J.L., Yadrick, R.M., Perrin, B.M., Knight, J.R., Eschenbrenner, A.J., Kueter, F.E., & Feldsott, S. (1988, October). Research and development of the training decisions system (AFHRL-TR-88-50). Brooks AFB, TX: Draft report prepared for the Training Systems Division, Air Force Human Resources Laboratory.

NEW TECHNOLOGIES FOR DEVELOPING AUTOMATED DATA-BASED
SPECIALTY KNOWLEDGE OUTLINES

Paul P. Stanley II
USAF Occupational Measurement Center
Randolph Air Force Base, Texas

Specialty Knowledge Tests (SKTs) are key to the success of the Air Force's enlisted promotion system. This panel provides a status report on research aimed at translating computerized occupational analysis data directly into draft test content outlines for use by teams of senior NCOs assigned to construct SKTs. The papers discuss the long history of attempts to make complex CODAP data more usable by test writers, AFHRL's procedures to screen and reformat the data, research findings which indicate that the automated procedures work, and the perceptions of the test psychologists who used the automated procedures to develop operational SKTs. The discussant commended the panelists on the balanced approach they had taken, through discussion both of the problems and the accomplishments involved in the research. He further observed that MTA conferences provide an ideal forum for this type of interchange, encouraging a unique interaction of researchers and interested outsiders which could not otherwise take place.

AUTOMATED TEST OUTLINE DEVELOPMENT RESEARCH FINDINGS

Authors: Weissmuller, Johnny J., The Texas MAXIMA Corporation
Dittmar, Martin J., Metrica, Inc.
Phalen, William J., Air Force Human Resources Laboratory

INTRODUCTION

This paper is a report of the research findings related to the production and use of automated test outlines (ATOs) for Air Force Specialty Knowledge Test (SKT) construction. After a short review of the process used to develop automated test outlines, the major focus will be on the reliability and validity of obtained results.

DEVELOPMENT PROCESS

Briefly, task-level predicted testing importance (PTI) values derived from off-the-shelf task factor components of Training Emphasis (TE), Task Learning Difficulty (TD), Percent Members Performing (PMP) and Percent Time Spent (PTS) by grade, and Average Grade Performing (AG) at E-5 and E-6/7 levels are used to delimit Air Force Specialty (AFS) knowledge domains in terms of restricted subsets of tasks from full task inventories. These subsets of tasks are then administered by mailout approximately 3 to 4 months prior to the start of SKT construction projects to random samples of senior NCOs (50 - 70) who currently work in the AFSs. These NCOs rate each task within the subset on a 7-point scale of specialty knowledge testing importance. These resulting field inputs are then processed, analyzed, and subsequently used to determine testing importance weights for each task in the mailout and to calculate test outline weights (numbers of test items to be written) for each major duty area of the specialty. These task-level testing importance ratings by field NCOs are key components within the automated test outline development process.

RELIABILITY AND VALIDITY ESTIMATES

Specialty Knowledge Tests are normally developed 6 to 18 months before their scheduled administration to Air Force enlisted populations. Consequently, direct reliability and validity estimates for those SKTs constructed from automated test outlines will not be available until after the test administration cycle is completed (early 1989). However, the "goodness" of the process used to generate automated test outlines can be evaluated by examining reliability and validity indices associated with a primary component of the process - field-validated testing importance (FVTI).

To date, automated test outlines have been developed for 28 Air Force Specialties. Table 1 lists interrater reliability estimates of FVTI for each of those specialties.

TABLE 1
FVTI INTERRATER RELIABILITY

Specialty		N-Tasks		N-Raters		R11		Rkk	
		E-5/6/7	E-5	E-6/7	E-5	E-6/7	E-5	E-6/7	
(In-Flight Refueling)	112X0	127	43	43	.478	.473	.975	.975	
(Survival Training)	121X0	148	31	30	.225	.199	.900	.882	
(Still Photography)	231X2	145	25	26	.168	.148	.835	.818	
(Audiovisual Production)	232X0	135	24	24	.295	.323	.909	.920	
(Safety)	241X0	162	45	45	.270	.282	.943	.946	
(Command & Control)	274X0	147	30	30	.294	.282	.926	.922	
(Aircraft Cont & Warning)	303X0	152	46	51	.117	.121	.859	.875	
(Space Systems Maint)	309X0	214	24	24	.123	.156	.757	.803	
(Defensive Fire Cont)	321X1E	145	6	6	.405	.397	.792	.787	
(Defensive Fire Cont)	321X1G	165	15	15	.384	.397	.903	.908	
(PMEL)	324X0	161	29	29	.233	.255	.897	.908	
(Maint Scheduling)	392X0	149	18	18	.376	.366	.916	.912	
(Aircraft Fuel)	423X3	155	24	25	.205	.208	.861	.868	
(Aircraft Pseudraulic)	423X4	179	27	28	.164	.223	.840	.889	
(Fabrication & Parachute)	427X3	175	40	41	.281	.229	.940	.924	
(Helicopter Mech)	431X0	187	39	41	.228	.228	.938	.942	
(Special Vech Maint)	472X0	257	30	30	.218	.249	.892	.907	
(Elect Power Lines)	542X1	198	25	25	.275	.298	.903	.914	
(Elect Power Production)	542X2	155	40	41	.188	.232	.902	.925	
(Structural)	552X0	202	33	34	.132	.138	.834	.845	
(Cost Analysis)	674X0	148	33	33	.192	.217	.887	.901	
(Social Actions)	734X0B	137	25	24	.245	.281	.887	.901	
(Combat Arms)	753X0	222	11	11	.316	.367	.836	.864	
(Public Affairs)	791X0	147	31	31	.207	.319	.890	.936	
(Radio & TV)	791X1	156	22	22	.293	.273	.901	.892	
(Environmental Medicine)	908X0	192	34	35	.247	.308	.917	.939	
(Mental Health)	914X0	140	25	24	.302	.263	.914	.894	
(Medical Material)	915X0	190	31	34	.192	.351	.880	.948	

The Table 1 data show reasonably good interrater reliability estimates (all R11=0 probabilities were less than .01).^{*} Over 75% of the R11s exceeded .20 and no Rkk estimate was below .75. Although these levels of reliability are in accordance with expectations, as judged by the Air Force's ongoing experiences with the reliability of field NCO ratings of task learning difficulty, future research efforts will be directed towards increasing interrater reliabilities, primarily through clearer and more simplified rating booklet instructions and the identification and segregation of more relevant rater subgroups. The low number of raters in some AFSCs was attributable to a variety of causes: small rater populations, Testing Importance Survey booklets returned too late for inclusion in ATO projects, high percentage of senior NCOs on TDY, leave, or PCS, and higher than average percentage of deviant raters.

Since the subject-matter experts (SMEs) who develop SKTs are from the same population as those selected to complete testing importance field rating booklets, we were able, in a limited number of cases, to readminister the rating booklets to assess the stability of FVTI ratings over time. Results of this test-retest effort are contained in Table 2.

* Deviant raters, whose correlations with the composite of all raters were not significantly better than zero at $\alpha = .05$, were eliminated from the calculation of R11 and Rkk (usually less than 10% of sample).

TABLE 2
TEST-RETEST RELIABILITY (FVTI)

AFSC	N-Raters	X-First Administration		Y-Second Administration		Avg $r_{x,y}$
		\bar{x}	sd	\bar{x}	sd	
42353	1	4.4	.61	4.3	.89	.54
42373	1	4.3	.56	4.5	.89	.58
39250	2	4.4	1.86	4.1	1.95	.77
39270	2	3.9	1.57	3.8	1.85	.66
90850	1	2.5	1.61	2.6	1.43	.79
90870	1	2.5	1.61	2.6	1.43	.79
54252	1	3.8	.99	4.2	.64	.45
54272	1	3.9	1.09	4.3	.68	.43
55250	1	5.2	1.13	4.9	.81	.58
55270	1	5.4	1.17	5.1	.86	.58
67450	2	4.3	1.23	4.2	.77	.67
67470	2	4.3	1.27	4.6	.95	.61
32151E	1	4.4	1.29	4.3	.93	.49
32171E	1	4.4	1.29	4.3	.93	.48
42753	1	4.7	.86	5.2	.64	.43
42773	1	3.9	1.64	4.2	.64	.27
32151G	?	4.7	1.40	4.0	1.28	.77
32171G	2	4.7	1.34	4.0	1.29	.76
12150	1	5.3	1.53	4.3	1.64	.61
12170	1	4.3	1.46	4.0	1.51	.39

In approximately 75% of the cases, test-retest correlations were .5 or higher. Given that there was a 3-to-4 month period between first (x) and second (y) administrations and totally different administration environments (x was self-administered at the rater's home station, and y was administered at the USAF Occupational Measurement Center by the contractor as part of the initial SKT construction inbriefing), these correlations in most instances tend to support rating stability. They are 1.5 to 3 times higher than the interrater agreement coefficients (Rlis) in Table 1, except for AFSCs 321X1E and 427X3. There was no discernible trend with respect to mean ratings between first and second administrations. Rater variance tended to be smaller for the second administration ($P < .02$, Sign-Rank Test of Differences) indicating a conservative rating policy which may be the result of the more structured second administration environment.

Although the validity of SKTs constructed from automated test outlines cannot at this time be assessed directly in terms of increased job relevance of test content, it can to some extent be inferred by examining characteristics of FVTI ratings. As previously stated, FVTI ratings are used to establish automated test outline weights (the recommended numbers of test items to write for each major duty area). For any given Air Force specialty the number of these major duty areas can vary from as low as 8 to as high as 26. It seems reasonable to assume that the extent to which SKT construction teams adhere to these recommended duty area weights is an indication of the SME-judged goodness (validity) of the FVTI ratings and, to a lesser extent, the SME-judged appropriateness of the automated test outline development process. Table 3 shows the correlation between recommended and final (adjusted by the test construction team) major duty area weights. An alternative explanation for the high degree of adherence to the ATO weights by the SMEs, in the opinion of several test psychologists, was the flexibility the SMEs had in selecting knowledge requirements when writing an item on a task.

TABLE 3
CORRELATIONS BETWEEN RECOMMENDED AND ACTUAL TEST OUTLINE WEIGHTS USED

AFSC	Number of Outline Areas	r	AFSC	Number of Outline Areas	r
42353	16	1.00	23152	10	.99
42373	16	1.00	23172	10	.98
79150	16	1.00	54252	23	1.00
79170	16	1.00	54272	23	1.00
73450B	12	.96	55250	23	1.00
73470B	12	.96	55270	23	1.00
42354	14	1.00	67450	12	.98
42374	14	1.00	67470	12	.95
79151	16	.99	54251	16	1.00
79171	16	.99	54271	16	1.00
91350	15	.98	32151E	12	.83
91570	15	.99	32171E	12	.84
39250	17	.96	42753	26	.99
39270	17	.95	42773	26	.99
90850	16	.85	11250	8	1.00
90870	16	.94	11270	8	1.00

The correlations listed in Table 3 range from a low of .83 (AFSC 32151E) to 1.00. For the total set of Air Force Specialties, approximately 88% of the automated outlines had correlations between recommended and actual weights of .95 or higher. In approximately 44% of the cases, no weight change was necessary. These are positive indications and speak to the validity of the FVTI ratings.

In addition to being used to calculate recommended major duty area weights, FVTI ratings are also used to differentiate outline tasks into A (high testing importance) through D (low testing importance) categories, depending on the mean FVTI value computed for each task. If field-validated testing importance ratings are valid, one would expect a greater percentage of A tasks to be used to generate test items than B, C, or D tasks; to a lesser extent, one would expect B tasks to be used to generate test items at a somewhat higher rate than C or D tasks, and that D tasks would have the lowest usage rate of all. Table 4 lists for each task category (A, B, and C) the ratio of the proportion of test items written on tasks in that category to the proportion of tasks in that category appearing in the E-5 and E-6/7 outlines combined. The D category is not listed, because only one D task was used by one AFS. Even though tasks in the D category were not to be used, except with written justification, it is never-the-less, significant that only one team felt the need to justify the use of only one D task. On the other hand, although SMEs were required to write a minimum of one item on each A task, it is significant that the item/task ratio for A tasks substantially exceeds 1.0 in all AFSs, and there are no reversals in the expected decrease in item/task ratios from A to B to C.

TABLE 4
TASK USE RATIO

AFS	Task Type	N (Tasks)	Item/Task Ratio	AFS	Task Type	N (Tasks)	Item/Task Ratio
423X4	A	11	1.9	552X0	A	53	2.1
	B	83	1.2		B	160	1.1
	C	209	.8		C	166	.4
915X0	A	11	1.3	674X0	A	60	1.3
	B	40	.7		B	97	.5
	C	189	.7		C	96	.4
392X0	A	48	1.8	542X1	A	70	1.9
	B	42	1.4		B	134	.8
	C	156	.6		C	129	.7
908X0	A	47	2.4	321X1E	A	64	2.0
	B	40	1.7		B	94	.4
	C	161	.5		C	94	.1
231X2	A	45	1.9	427X3	A	54	2.6
	B	111	1.2		B	133	.8
	C	112	.4		C	139	.5
542X2	A	53	1.1	112X0	A	38	1.9
	B	97	.7		B	106	.9
	C	97	.6		C	106	.3

Tables 5 and 6 examine the relationships between FVTI and five task-level factors: Predicted Testing Importance (PTI), Percent Time Spent by Members Performing (PTM), Training Emphasis (TE), Task Learning Difficulty (TD), and Average Grade Performing (AG). Because of the large number of tasks used to compute these correlations, a coefficient of $\pm .17$ is significant at $\alpha = .05$ (two-tailed), and $\pm .22$ at $\alpha = .01$ (two tailed).

TABLE 5
CORRELATION OF FVTI (E-5) WITH OTHER TASK FACTORS

AFS	PTI (E-5)	PTM (E-5)	TE	TD	AG	N (Tasks) E-5/6/7
112X0	.55	.28	.72	.49	.00	127
121X0	.69	.13	.81	.34	-.37	148
231X2	.43	.47	.53	.09	-.11	145
309X0	.22	.00	.12	.57	-.03	214
321X1E	.51	.40	.49	.58	-.15	145
321X1G	.80	.44	.73	.65	-.25	165
324X0	.55	-.04	.41	.55	-.01	161
427X3	.72	-.16	.61	.57	-.38	175
542X1	.83	-.21	.65	.72	.07	198
542X2	.41	-.18	.18	.66	-.02	155
552X0	.53	.07	.71	.08	-.02	202
674X0	.64	-.08	.64	.40	.25	148

TABLE 6
CORRELATION OF FVTI (E-6/7) WITH OTHER TASK FACTORS

AFS	PTI (E-6/7)	PTM (E-6/7)	TE	TD	AG	FVTI (E-5 vs E-6/7)
112X0	.37	.38	.58	.61	.21	.97
121X0	.54	.21	.65	.43	-.23	.93
231X2	.33	.27	.29	.34	.14	.89
309X0	.18	.08	-.04	.50	.25	.92
321X1E	.57	.13	.46	.58	-.11	.99
321X1G	.80	.25	.67	.71	-.12	.98
324X0	.54	-.16	.11	.74	.30	.89
427X3	.69	-.11	.40	.65	-.10	.94
542X1	.81	.29	.48	.73	.32	.94
542X2	.45	.05	-.02	.74	.30	.91
552X0	.49	.19	.55	.12	.11	.92
674X0	.68	-.07	.55	.43	.35	.95

As can be seen in Tables 5 and 6, FVTI correlations with PTI at both the E-5 and E-6/7 levels are relatively high, the single exception being AFS 309X0. This AFS is probably the most diverse (heterogeneous) of all those for which outlines were developed. This diversity may account to some extent for the relatively weak relationship between FVTI and PTI. The relationships among FVTI, TE and TD are also in the expected direction. At the E-5 level we would expect TE to have a stronger impact on FVTI than at the E-6/7 level, since TE is essentially a measure of recommended training emphasis for first-term airmen. Conversely, we would expect TD to have a stronger relationship with FVTI at the E-6/7 level than at the E-5 level, since TD is an estimate of how difficult it is to learn to perform a task. Both of these expectations are confirmed by the correlations in Tables 5 and 6, which lend a degree of convergent validity to the testing importance measure. On the other hand, the consistently strong, positive relationship between FVTI (E-5) and FVTI (E-6/7) could indicate the presence of an unwanted autocorrelation resulting from the dual-column "E-5/E-6/7" format of the rating booklets used to collect FVTI information.

It is evident from these findings that a single PTI equation to predict FVTI is not a satisfactory procedure, and that more attention must be given to TD, which has been underweighted in the procedure for selecting tasks for the Testing Importance Survey.

CONCLUSION

Although the statistical information gathered thus far is by no means overwhelming, it is very encouraging that it is uniformly in the right direction for almost all AFSs in which the occupational data-based automated outline procedure has been applied. From a validity standpoint, the most telling evidence is yet to come. Final judgments must wait until the statistical characteristics of the end products (the administered SKTs) are analyzed and, most importantly, until comments from SMEs on subsequent revisions of ATO-Developed SKTs and from Air Force examinees can be assessed. Examinee comments will come from two sources: a brief survey administered to the examinee before he leaves the testing room and complaint letters sent to the USAF Occupational Measurement Center (USAFOMC). The expectation is that there will be significantly fewer comments than in the past regarding lack of job relatedness of test items.

Development of Automated Data-Based
Specialty Knowledge Test Outlines: Current Procedures

1Lt Kathleen M. Longmire, Air Force Human Resources Laboratory
William J. Phalen, Air Force Human Resources Laboratory
Johnny J. Weissmuller, The Texas MAXIMA Corporation
Martin J. Dittmar, Metrica, Inc.

Within the Air Force, promotion to SSgt (E-5) through MSgt (E-7) enlisted grades in most Air Force Specialties (AFSS) is determined by an airman's relative rankings in the Weighted Airman Promotion System (WAPS). Components of the WAPS include the individual's Airman Performance Reports (APRs), Time in Grade and Service, as well as scores on the Promotion Fitness Examination (PFE) and the Specialty Knowledge Test (SKT). Often the most critical factors in determining promotion success are the PFE and the SKT. The SKT is a 100-item multiple choice test designed to measure job knowledge in Air Force job specialties. The Test Development Division (OTL) within the USAF Occupational Measurement Center (USAFOMC) is responsible for the development of SKTs and has maintained a very successful program in this area since the 1950s.

Also within the USAFOMC is the Occupational Analysis Division (OAY), responsible for conducting occupational surveys for most enlisted career specialties. As part of this program, detailed task inventories containing as many as 2000 tasks are developed for each career field. Content of these inventories is based on input from experienced senior-level NCOs and on direct observations made by inventory developers on a variety of enlisted personnel performing their jobs at various duty locations. Once developed, job inventories are mailed to a representative sample of hundreds, or even thousands, of individuals within each AFS. Survey respondents are asked, first, to check which tasks they perform and then to rate the relative amount of time they spend performing those tasks. This same job inventory is also sent to a sample of 50-200 senior NCOs, who rate each task on the emphasis it should receive in formal training programs for first-term airman and how difficult it is to learn to perform the task. From these ratings, task-level data on Percent Members Performing (PMP), Percent Time Spent (PTS), recommended Training Emphasis (TE), and relative Task Difficulty (TD) are obtained for each individual task in the inventory.

To enhance the content validity of the SKTs, USAFOMC has actively pursued the development, implementation, and refinement of procedures for integrating the occupational survey report (OSR) data into its test development procedures. The goal of the integration process has been to formalize an operationally feasible SKT development methodology that ensures SKTs more directly satisfy the job-relatedness criterion of the Uniform Guidelines on Employee Selection Procedures. The intent is to be able to validate the promotion tests against a systematic and appropriately comprehensive occupational analysis of the career field for which the SKTs are to be used. By 1984, after a decade of effort, USAFOMC realized that further integration of occupational analysis data into the test development process would require the direct intervention of these data into the core of the test development process, namely the development and weighting of the test outline itself. Such an undertaking would require considerable research to develop a valid, reliable, and operationally feasible procedure for producing a data-based outline. That year, USAFOMC formally requested AFHRL to undertake the necessary research.

This paper deals with the automated procedures developed by AFHKL to generate test outlines which allow SKI developers to link performance-based task statements with knowledge-based test items. It first describes current methodology used to generate automated test outlines (ATOs), including procedures for predicting which job inventory tasks are most relevant for testing purposes; methods for combining these tasks into a survey instrument for collecting direct measures of task-level testing importance (TI); and finally, procedures, based on average TI values, for specifying the number of test items to be written at each level of the automated outline. The paper then discusses possible modifications which could enhance the outline's content validity and acceptance by subject-matter experts (SMEs). These modifications include ways to validate and update tasks in the TI surveys and alternative procedures for incorporating semantic and task clustering techniques into the outline development process.

CURRENT PROCEDURES

Conventional Test Outline (CTO) Development

Promotion tests within the Air Force are developed by teams of two to eight SMEs, usually Master Sergeant or higher, who are selected by their representative major commands. Under the direction of a trained test psychologist, who acts as group facilitator and ensures that team members adhere to USAFOMC guidelines, these teams write test questions designed to assess the extent to which airmen possess the job knowledges needed to perform the tasks which comprise their AFS. To ensure a representative sampling of the knowledges required within an AFS at different skill levels, team members draw not only on their own knowledge and experience, but also from information within Air Force Regulation 39-1, Airman Classification, and from the most recent Specialty Training Standard (STS), which outlines all functions within an AFS for the purpose of training and evaluating personnel in those skills. Teams also have access to data from the most recent occupational analysis study for their AFS. They are provided with computerized lists of analysis data for each duty and task. These lists include data on PMP and PIS, and also show the relative TE and LE ratings supplied by senior-level MCOs. Typically, SMEs have found these data difficult to use since most have little or no expertise in interpreting occupational survey reports. These reports often contain large volumes of information and are in formats not readily applicable for test development purposes. Recency of the data is also a concern with both SMEs and test psychologists who often view data over 2-3 years old as obsolete.

Despite some difficulties, it is widely accepted that the use of these data should increase the tests' validity. Although the MCO test developers represent a wealth of knowledge and have years of personal experience in their career fields, their small number (typically 2-8 SMEs on a team) raises questions about the representativeness of their experience across the entire career field which is being tested. Because survey data are gathered from such a large number of job incumbents from each AFS, they should provide a more representative sampling of task performance than could possibly be provided by a small number of even the most experienced and knowledgeable SMEs.

Under existing procedures, JET major revision teams meet every two years to perform a thorough rewrite of the tests. SMEs spend the first 1-2 days of their revision team meeting on alternate years to update the previous year's major revision of the tests.

the test development process generating and weighting the outlines for both the E-5 and E-6/7 promotion tests. Using the SIS as their base document, they identify the major knowledge areas necessary to perform tasks within their AFS, and hence, define the knowledge components of their test items. Using an iterative process, these general knowledge areas are broken down further into more and more discrete subdivisions. Once knowledge components have been assigned at their lowest level, SMEs next allocate percentage weights to each major area of the outline and then distribute these weights accordingly among the various subparts. Using new procedures developed by AFHKL, which owe a great deal to feedback from SKI development teams, the CIO development process has been replaced by generation of an automated test outline (ATO) delivered to USAFOMC prior to the team's arrival. It is this outline which will be the focus of the remainder of the paper.

Automated Test Outline (ATO) Development

The first step in the ATO process is to predict that subset of tasks from the USAF Job Inventory which have the highest relevance for testing purposes. The goal is to select between 150-200 tasks which will later be validated by senior NCOs in the field. Using a standard weight regression equation which considers PMP, PTS, TE and TD values for each task, a predicted TI (PTI) value for each task is computed. PTI values for E-5 and E-6/7 tasks are computed separately, based on the separate E-5 and E-6/7 composite job descriptions computed from occupational survey data. To reach the minimum requirement for 150 tasks, the top 120 items, based on their PTI values from both the E-5 and the E-6/7 lists, are first selected. Added to this list are any tasks performed by at least 50% of either the E-5 or the E-6/7 groups or any tasks which have a TE value at least 1 standard deviation (S.D.) above the mean for TE.

The second step in the process is to combine these 150-200 tasks under their appropriate OSR duty headings to create a Testing Importance Survey booklet used to gather expert opinion regarding the TI of each task. TI is defined as how important it is to include, in an SKI, the job knowledges needed to perform that task. A task should have high TI if it requires knowledges that are critical to successful performance within each AFS. Approximately 2 months prior to the major revision team's arrival, a sample of between 50-70 senior NCOs (E-7/8/9s) is selected and mailed a booklet containing a brief background information section and a section containing the tasks selected for their AFS. Respondents are asked to rate each task on a 1-7 Testing Importance scale, where "1" indicates "No Importance," and "7" indicates "Extremely High Importance." Respondents provide only one rating for each task if they believe that the level of TI is the same for the E-5 and the E-6/7 SKI. However, they are instructed to provide two separate ratings if they believe the TI value is different for the grade levels.

Once booklets are returned, the data are entered into a PC-system and are processed for interrater reliability. Deviant raters, whose correlations with the composite of all raters is not significantly better than zero at $p=.05$, are eliminated from the computation of interrater reliability. Based on all remaining raters (usually greater than 90%), an average testing importance value is computed for each task. This process is accomplished twice, once for the E-5s and a second time for the E-6/7s. Mean task-level TI values are squared to increase the relative value of high TI means and then summed within each major duty area to obtain a Duty Sum of Squares (DSS) for each duty area of the E-5 and E-6/7 ATOs. These duty sums, divided by the Total Sum of Squares (TSS) across all duties, are the numbers which determine the duty area

weights in terms of number of test items apportioned to each. Two different numbers are computed in the process: the numbers of WRITE ITEMS and KEEP ITEMS. The number of WRITE ITEMS is the total number of test items to be written, including the selected and alternate test items; the number of KEEP ITEMS is the number of selected items which will become part of the 100-item SKT.

Within each duty area, tasks are assigned a letter value (A,B,C,D) indicating their relative testing importance. "A" tasks, assessed to be of extremely high importance, are those tasks which have average TI values at least one S.D. above the grand mean or which have a value of 6.0 or above (on the 1-7 scale). "D" tasks, determined to be the least important for testing purposes, are those which have average TI values at least 1 S.D. below the grand mean or which have a scale value of less than 2.0. The upper 50% of the remaining tasks are designated "B" tasks and the lower 50% "C" tasks, although any task with a scale value of 4.0 or above must be at least a "C."

Ideally, more test items should be written on tasks with "A" values than on tasks with "B" values, and more test items should be written on tasks with "B" values than on tasks with "C" values. Tasks with "D" values should generally be ignored. Teams are encouraged to focus their item writing efforts accordingly. Under present ATO test development rules, a maximum of three test items can be written against any one task statement. Specific justification and approval to exceed this level are required. At least one test item must be written against each task statement assessed to be of extremely high testing importance ("A" tasks), and again, specific justification is required for failure to do so. Where the number of "A", "B", and "C" tasks is insufficient to generate the required number of WRITE ITEMS and KEEP ITEMS within any major duty area, "D" tasks may be used with the review psychologist's approval.

The key to the success of the ATO lies in the commitment of the test developers to write test items based on the important knowledge components underlying the performance task statements that comprise the test outline. In a number of cases, this association is relatively easy to form. In others, however, it requires effort and analysis on the parts of both the test psychologist and team members to write test items. This analysis should start with a determination of knowledges required to successfully perform the task. The scope of this analysis can then be expanded to knowledges related to the context within which the task is performed. These include knowledges associated with determining "why," "when," "where," "if," and "how" to perform the task. These defined knowledge components can then be used individually or in combination to form the content area from which the test is written.

FUTURE DIRECTIONS AND MODIFICATIONS

One of the more salient criticisms of the ATO has been the complaint that tasks requiring the same knowledge are scattered throughout the outline and that tasks grouped under the same duty heading require quite different knowledges. This criticism, of course, points to a problem in the organization of the outline, rather than to a deficiency in its content. One effort to alleviate this problem involves the use of micro-based software to permit teams to modify the ATO interactively. The micro-based procedures would permit a team to impose a more workable organization on the set of tasks

in the ATO without having to make all the necessary changes manually. The software would reorder tasks, recalculate weights, and print a revised outline document based on input from SMEs.

Another promising approach to improving the organization of the outline is through the application of semantic and co-performance task clustering procedures for grouping the tasks. Semantic clustering may be used to examine the semantic content of tasks, relate tasks to other tasks with similar content, or link tasks by relating them to specific components of a common semantic data base. Examples of such data bases would be the word processing data files that contain the career development course (CDC) materials or technical manuals to which test items supporting the tasks in the ATO have been referenced. On the other hand, task clustering would allow the grouping of tasks into discrete modules based on the degree of task co-performance. Tasks within a module should share similar underlying knowledge components. By including at least one task, either the most representative or the most discriminating, from within each cluster in the 11 survey, a more representative sampling of knowledge areas could be achieved.

A second difficulty with the ATO procedure has been that it relieves the team of its former responsibility to think about, discuss, and agree on the content of the test. Handing the team an already prepared outline deprives the members of a valuable learning and communication activity and of a feeling of ownership and commitment to the product. One remedy for this situation would also help solve the outline organization problem previously mentioned. As under traditional procedures, team members would first participate in an interactive group process of discussing, defining, and agreeing on the important knowledges which should comprise the test. Their second step would be to map each ATO task statement under its appropriate knowledge area. For the final outline, tasks could be reordered so as to align tasks for which the same knowledge area had been indicated. As stated, this approach will encourage the desired team interactions and should also result in a better task-to-knowledge mapping within the outline.

A third problem with the ATO, identified especially when working with a complex AFS, is that test development teams may believe important tasks have been omitted. Either important tasks were dropped by the screening process or they were not included in the original job inventory because they are associated with new equipment, new procedures, etc. within the AFS. Input from minor revision teams may prove extremely valuable in addressing this concern. Minor revision teams, normally containing only 2 members, meet on the off-years for a period of 11 days to review test items for compromise, keying errors, etc., and to replace at least 10% of the test items. Their assistance could be used in validating and updating the task list which ultimately becomes the basis of the ATO.

At present, a single algorithm is used to predict those tasks which have high testing importance. Research is already underway to determine whether a single prediction equation is sufficient for all AFSs, or whether the wide diversity of AFSs tested requires several differentially weighted equations to ensure that relevant tasks are not omitted from the 11 survey. As an additional safeguard, input from knowledgeable SMEs would be useful in verifying that such omissions have not occurred. For example, the standard weight equation automatically deletes any tasks which fall under the A-D duty areas, normally designated for supervisory, nontechnical tasks. Knowledges associated with these tasks are more appropriately tested by the Promotion

Fitness Examination (PFE) and the USAF Supervisory Examination and, therefore, should not be included on an SKI. However, this assumption may be causing important job knowledges (e.g., administrative or clerical tasks) to be missed. The problem is currently being remedied to some extent by computing a CODAP group difference description between composite job descriptions for the lowest three grades (E-1, E-2, and E-3) and the highest three grades (E-7, E-8, and E-9). Those tasks from Duties A, B, C, or D which show a positive difference in favor of the lower grade group are considered for inclusion in the Testing Importance Survey.

A better approach may be to use the minor revision teams to confirm the task selections made by use of the group difference description and to identify additional tasks missed by the computerized procedure. To assist SMEs in validating the screening process, they would be provided with three lists of tasks that were excluded from the proposed subset of tasks. One list would contain those technical tasks for which PMP for E-5s or E-6/7s falls between 30% and 50% or for which TE is at least average. A second list would contain those nontechnical tasks (those in duties A-D) for which PMP is equal to or greater than 30% or for which TE is at least average. A third list would show all the remaining tasks in the inventory.

SMEs will be asked to review these lists in order to determine if any of the tasks involve omitted knowledges that are important for testing on an SKI and, as a result, should be added to the proposed list of tasks. Reasons for including the task would be annotated. Team members would next be asked to identify emerging areas of work requiring new knowledges that should appear on the SKI. Here, SMEs will be asked to review the original list to determine if new or emerging tasks should be added to any existing duty. If so, they are instructed in how to develop acceptable statements for each "new" task. Likewise, they identify any new duty areas, and again, using task construction guidelines, they develop supporting task statements for each duty. In addition to new and emerging areas, the minor revision teams will identify obsolete, redundant, and poorly worded tasks. This information will be passed on to USAFOMC/OMY to assist them in deciding when to revise an inventory and resurvey the AFS.

We still have much to learn about various kinds of complex SKI projects as, for example, those involving multiple career ladders that are divided at the 3- and 5-skill level, but joined at the 7-skill level; or those which have shredout sub-specialties within a career ladder; or an AFS, such as the Precision Measuring Electronic Equipment (PMELE) repair specialty, which is highly fragmented in its responsibility for a large array of diverse pieces of equipment and has very few tasks that are common to everyone in the specialty. Producing valid outlines in the face of these complexities will be a major part of our goal during the coming year. We still have much to learn. However, the procedures and potential improvements reported here will go far in improving our ability to generate valid, operationally acceptable SKI outlines. The combined, cooperative efforts of the USAFOMC and AFHRL are needed to see this project through to successful completion and operational implementation. Its importance to the promotion and retention of worthy members of our enlisted force cannot be overemphasized.

Automated Specialty Knowledge Test Outline Procedures:
A Management Perspective

Paul P. Stanley II
Ronald C. Baker
Joseph S. Tartell

USAF Occupational Measurement Center (USAFOMC)

Specialty Knowledge Tests (SKTs) play a key role in US Air Force enlisted promotions. In 1970, partly to facilitate development of these important tests, the Air Force combined the operational portion of its occupational analysis program with the organization responsible for SKT development, forming what is now called the USAF Occupational Measurement Center (USAFOMC). In theory, promotion tests were, from that point on, to be based on the data being collected to make personnel utilization and training decisions using the Comprehensive Occupational Data Analysis Programs (CODAP).

That, however, is not what happened. Instead of the desired synergistic effect of colocating these two programs so that each might foster improvements in the other, the organizations continued to work independently of, and sometimes antagonistically toward, one another. To the psychologists and subject-matter experts (SMEs) who developed SKTs, CODAP-based occupational analysis products were not practical for use. The information typically consisted of hundreds of pages of task data tables, intimidating in complexity and discouraging in volume. The Air Force's CODAP approach to job analysis was viewed by many test developers as new and unproven, and not necessarily an improvement to the already quite successful test development procedures which had evolved in the 1950s and 1960s. In short, to managers on the test development side of the house in USAFOMC's predecessor organization, CODAP-based occupational analysis was viewed as the solution to a problem which did not exist. Thus, valuable information which could enhance validity was not being adequately used.

Importance of SKTs

SKTs are 100-question multiple-choice written tests. They are written at USAFOMC by teams of senior NCOs from the field, who provide the subject-matter expertise, and USAFOMC test development psychologists, who provide the test-writing expertise. SKTs are revised annually, to protect against compromise and the effects of procedural and technological change. The tests undergo a "major" revision one year, with less extensive "minor" revisions on alternate years. Major revision teams develop a test content outline before writing questions. Minor revision teams generally revalidate and use the previous major team's outline.

SKTs are critical to the success of the Air Force's Weighted Airman Promotion System (WAPS). Under WAPS, airmen compete for promotion to the ranks of staff sergeant (E-5) through master sergeant (E-7) with other airmen in the same Air Force specialty (AFS) on the basis of a single score. This single WAPS score is the sum of six component measures, with the SKT accounting for up to 22% of the total. Because most of the other factors do little to disperse promotion competitors, SKTs are often the deciding factor in determining who gets promoted.

Conventional Test Outline Development

The test content outline is developed during the first week of a project, after team members review the old tests and become familiar with the control documents, in particular the specialty training standard (STS). The STS is an official training document which lists the major duties and responsibilities of

members in an AFS, and which is available to all managers and members in that specialty. The essence of the outlining process is to divide specialty knowledge into four to seven categories which are all-inclusive but mutually exclusive. The SMEs negotiate among themselves the relative weights for the areas they've identified, and then add subcategories so that each outline area consists of a manageable number of test questions--seven questions is the upper limit imposed at USAFOMC. Occupational analysis data are used to the degree felt appropriate by members of the test development team. Team members are encouraged to use them, but requirements for data use are loosely stated and commonly circumvented by team members who are uncomfortable with the data. This may be viewed as a shortcoming, but the intent is to ensure that those in the field know that USAFOMC tests are developed "by airmen for airmen," and it is the SMEs, in the end, who must stand behind the weights assigned.

Need for Job Analysis

Because SKTs are so visibly important, a great deal of attention has been paid to maximizing their validity by incorporating the guidance of leaders in the field of test development. This has been especially true since 1978, when the *Uniform Guidelines on Employee Selection Procedures* were issued. It is a USAFOMC goal to comply with the *Guidelines*, though the Air Force is exempt from them for military personnel decisions. The *Guidelines* have been criticized for how difficult it is to comply with their documentation requirements, but they represent a coordinated point of view of leaders in the professional testing community. Another authoritative source of guidance concerning validation strategies is the *Standards for Educational and Psychological Tests* (1984), produced jointly by the American Psychological Association, the American Educational Research Association, and the National Council on Measurement in Education. A third source of guidance incorporated in USAFOMC procedures is the *Principles for the Validation and Use of Personnel Selection Procedures* issued by the APA's Division of Industrial-Organizational Psychology (1980).

The common thread running through the *Guidelines*, the *Standards*, and the *Principles*, as well as through the different types of validity strategies covered in each, is the requirement that test content be clearly based on a thorough job analysis. Such an analysis becomes the basic source document supporting the appropriateness of test content. In the eyes of many, such analyses already existed which could be readily applied to the SKT outline development process: the CODAP-based occupational analyses of USAFOMC's Occupational Analysis Division. These task-based studies appear to meet most of the job analysis requirements recommended in the *Guidelines* and the *Standards*.

Despite their ready availability for many SKTs, the CODAP-based studies were not used to the extent that they could have been. One point of view was that content validity had already been established, and that the procedures and control documents used ensure that validity is built into the tests from the outset. Over the years, however, management thinking has evolved to the viewpoint that important information existed which was not being used. The less the data were used, the more individual test developers relied solely on their own experience, to the exclusion of using data describing the job experiences of the hundreds or thousands of incumbents who respond to occupational surveys. The ineffective or insufficient use of occupational analysis data is now viewed as a problem.

Early Efforts to Foster Use of CODAP Data

Numerous steps were taken over the years to make CODAP data more acceptable to test developers, including several attempts to tailor existing reports. "SKT

extracts" of CODAP data were designed which focused more specifically on the information which would be most useful to test developers. New scales were added by which to compare the relative importance of tasks. Tables were simplified as much as possible. Test developers were given specialized training in how to use the data, with emphasis on cutting through to essentials for the benefit of the SMEs. Still, there was reluctance to use the data.

Problems With the Available CODAP Data

CODAP data really were hard to work with, especially since the primary users were senior NCOs from the field with no experience in applying the results of occupational analysis. In a survey of USAFOMC test psychologists, Williams (1988) identified the following perceptions concerning the CODAP SKT extracts required to be used in test development:

- 1) They were too voluminous and complex: too many tasks, too many tables, too many scales.
- 2) They were too old: anything over 3 years of age was automatically considered "obsolete" by SMEs.
- 3) They contained too much information not relevant to testing: obviously oriented toward training and classification issues, or concerned with tasks which were not appropriate SKT content.

These types of perceptions had earlier led researchers to investigate the benefit which might be gained by automating the use of CODAP data. That is, since it was exceedingly difficult for the typical SMEs to make sense of the data because they viewed so much of it as extraneous, perhaps a computer might be programmed to do what a group of SMEs who had been well-trained and diligent in their approach would have done themselves.

Automated Test Outline (ATO) Research

As reported at the 28th Annual Conference of the Military Testing Association (Longmire, Phalen, and Weissmuller, 1986), research during the years 1975-1978 was directed toward maximizing the use of occupational analysis data in the test development process. Vaughan (1976) showed the feasibility of an automated approach to outline development. He derived regression equations from occupational analysis data and SME ratings of testing importance which reliably predicted the numbers of test items assigned to test outline areas by teams using conventional techniques. In the two Air Force specialties (AFSs) he studied, he used various combinations of variables to predict outline area weights produced by previous teams, finding the best single predictor to be the SMEs' ratings of the "importance" of the outline area. In subsequent studies, however, predictions did not fare as well, and it was observed that different equations might be necessary for different AFSs and even for different grades within an AFS.

Coincident to Vaughan's work was the research of Phalen (1978), who showed that SMEs' perceptions of task importance for testing can be captured, and an algorithm developed to describe the process. One of the most important aspects of his research was to show that the data could be made more manageable by eliminating nonrelevant information. Using SME ratings of task "usability" in SKT content, he developed "filters," which he then used to screen out nonrelevant tasks. He found that over half of the tasks in the original inventory could be eliminated. On the remaining tasks, he used four predictor variables (percent members performing, task difficulty, consequences of inadequate performance, and time delay tolerance) to obtain testing importance values for tasks. SMEs ranked the relative importance of each of these task

factors to testing importance. The result was a composite of task factors used to define testing importance for each task, which could then be used with occupational data for constructing and weighting an SKT outline.

Bills (1978) applied Phalen's methodology to four AFSs to show that the automated approach could be used in the operational setting. Using conventionally developed outlines as a criterion reference, Bills ran a comparative study which showed that the test content weightings prescribed by computer-derived outlines were not significantly different from those assigned by SMEs using conventional procedures. The Bills computer-derived outline was based on the STS. Each team was provided the most current STS and a printout in which tasks had been mapped into the subject areas of the STS. The SMEs found this format for their computer-derived outlines acceptable, and six of the eight reported that they preferred the new procedure.

Although the research of Vaughan, Phalen, and Bills produced promising results, only limited application followed, primarily because of the tangential problems involved. Most of the work was being done by psychologists whose primary job was operational SKT development. In addition, there was a problem with deviating from normal procedures for SKT development: SKTs are all operational promotion tests; it was feared that the influence of experimental procedures upon SME judgement would be perceived as being contrary to the aforementioned policy of "tests written by airmen for airmen." Automation of the outline process was considered, at best, to be an enhancement of an already well-accepted way of doing business, and was thus not readily accepted.

In 1984, at the request of USAFOMC, the Air Force Human Resources Laboratory initiated a contract for additional research. A primary goal was to create synthetically a task-level "testing importance" value primarily from off-the-shelf task factor data. It was felt that this procedure would have the beneficial effects of both demonstrating the job-relatedness of test items and avoiding the cost of USAFOMC routinely collecting an additional task factor. Research was proposed to collect direct ratings of testing importance and determine, first, if they were reliable, and, second, if the results obtained directly could be approximated by manipulating data already available. This was the basis of the present ATO research project, though the thrust has changed considerably.

ATO Development Procedures

With the procedures used in the present research, most of the groundwork is laid before the team of SMEs arrives at USAFOMC for the test development project. The first step is the generation of a Testing Importance (TI) survey from the original occupational inventory: the occupational inventory is screened using an empirically based algorithm which deletes material which is not relevant to test development. Typically, the screening process trims an inventory down from over 1,000 tasks to 150-200. This helps to eliminate the spurious influence of data on tasks which will not be covered on the test.

The condensed inventory is sent to a sample of 50-70 senior NCOs with instructions to rate each task's relative importance for promotion testing. The returned surveys are computer-analyzed, and each task given an importance rating (A, B, C, or D). This information is translated into a recommended number of questions to be written for each "duty area" into which related tasks have been grouped. This constitutes the test content outline if the team agrees. The team has the final say in test content, but must provide written justification for deviations from suggested weights.

ATO Advantages from the Management Perspective

The ATO approach to outline development successfully addresses the three biggest problems found with using CODAP-based occupational data:

- 1) The volume of the data reviewed is now manageable.
- 2) The TI survey provides an update of old data.
- 3) Nonrelevant information has been screened out.

From the standpoint of validity, the ATO approach provides direct documentation of the link between test questions and the specific tasks which have been determined to be most important for promotion selection. Even the most experienced team of SMEs can benefit from the collective judgments of the hundreds or thousands of job incumbents involved in a typical occupational analysis. From the standpoint of utility, its primary advantage is that it screens out the extraneous information that has made traditional extracts of occupational analysis data so unpopular among SMEs and test psychologists.

Ironically, one of the main goals of the research was attained but will be ignored. It was found that outline weightings could be generated from existing data, so that the expense of an additional task factor survey, for testing importance, could be avoided. However, it was also found that working with data collected only several months before ATO generation boosts the confidence of the SMEs in the data, even when the original study is several years old. For this reason, the Testing Importance survey will be included when ATO procedures are implemented. Since the new procedures will also include having the SMEs review and update the occupational inventory itself, USAFOMC's Occupational Analysis Division will benefit from the process as well.

Problems with the ATO

Phalen had earlier identified as a weakness the fact that CODAP task data do not specify the kind or degree of knowledge required for successful performance. This problem remains. Since any important job-related task may be based on a large number of important component knowledges, the SMEs have a great deal of leeway in writing about what appears to be a fairly narrowly defined topic, the task. However, that same "drawback" already holds true for subject-matter-based conventional test outlines. The idea with an ATO is that one can at least track back on any question and find a task-based reason (documented) for its existence. The burden is on the SMEs to ensure job-relatedness. This is an accepted practice: the courts in labor law give great deference to the judgment of subject-matter experts and supervisors as to what is important in a job. At the same time, it is the job of the test development professional involved in a project to continually remind the SMEs of this responsibility.

Another, and unexpected, problem was that an important part of the group dynamics process was removed and not replaced with the ATO procedures. The test outline development process was one during which the SMEs and test development personnel got to know one another and, in the negotiating process, develop a feeling of ownership of the outline and, thus, the test itself.

Future ATO Research

It was originally felt that automating test outline development would be fairly simple and straightforward, but that has not been the case. Subtle but nagging problems have hindered adoption of operational ATO procedures. As many and varied as the problems are, however, the basic premise remains the same: the concept of using automation to increase the acceptability of CODAP data to test

developers will increase the validity of the tests. For this reason, USAFOMC management is committed to continuing the project. Future research will be dedicated not to determining whether to implement ATO procedures, but, rather, how best to do so.

One of the concerns with the ATO approach, the group dynamics problem, may be addressed by redefining procedures so that the SMEs can work interactively with the data on a microcomputer, rather than be delivered a mainframe generated draft outline in advance. This approach was originally proposed by the contractor, but was turned down at that time by USAFOMC management because of a problem funding the microcomputers. In addition, it was felt that the basic ATO procedures would yield a product which was acceptable to a team of SMEs, as long as they had ultimate override authority in determining test content. This turned out not to be the case, as even teams whose members were in basic agreement with the task importance data wanted to reorganize the data into categories of their own making. For this reason, microcomputers will be a main feature of future research on this topic. We are confident that full-scale implementation of new procedures will occur in little more than a year.

References

American Psychological Association, Division of Industrial-Organizational Psychology. *Principles for the validation and use of personnel selection procedures*. Berkeley, CA: 1980.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education. *Standards for educational and psychological tests*. Washington: American Psychological Association, 1984.

Bills, C. G. Evaluation of computer-derived test outlines using conventional test outlines as a criterion reference during test development projects. *Proceedings of the 20th Annual Conference of the Military Testing Association*, 1978, 976-993.

Longmire, K. M., Phalen, W. J., and Weissmuller, J. J. Developing outlines for specialty knowledge tests using occupational survey data. Presentation (unpublished) at the 28th Annual Conference of the Military Testing Association, 1986.

Phalen, W. J. The development of a technique for using occupational survey data to construct and weight computer-derived test outlines for Air Force Specialty Knowledge Tests. *Proceedings of the 20th Annual Conference of the Military Testing Association*, 1978, 949-975.

Uniform Guidelines on Employee Selection Procedures. *Federal Register*, Volume 43, No. 166; August 25, 1978.

Vaughan, D. S. Prediction of test outline weights from occupational survey data. *Proceedings of the 18th Annual Conference of the Military Testing Association*, 1976, 435-460.

Williams, J. E. Results of the OSR SKT extract questionnaire given to test psychologists. Unpublished USAFOMC/OMD staff study, March 29, 1988.

Automated Specialty Knowledge Test Outline Procedures.
A Development Team Perspective

John E. Williams, 1Lt, USAF
Wendy L. Sotello, 1Lt, USAF
Paul P. Stanley II

USAF Occupational Measurement Center (USAFOMC)

Development of valid test content outlines is critical to the success of the Specialty Knowledge Tests (SKTs) developed at USAFOMC. The Air Force Human Resources Laboratory has developed a method to generate test outlines directly from occupational analysis data. This paper compares the new method for outline development with the conventional method currently in use at USAFOMC by examining the attitudes of test psychologists who have used both approaches.

Conventional Test Outline (CTO) Development

Every promotion test developed at USAFOMC is developed from an outline that determines test content and format. Subject-matter experts (SMEs) assigned TDY to USAFOMC normally construct the CTO early in the project, before writing any questions. SMEs use three primary documents in CTO development. The first and most important control document is the specialty training standard (STS). The STS is a comprehensive listing of important duties and tasks agreed to by career field managers. The second control document is the specialty description in Air Force Regulation 39-1, Airman Classification. This document describes all duties, responsibilities, and requirements in a given career field. The STS and specialty description define specific content to be covered by a given promotion test. The third document used is the SKT extract prepared by the USAFOMC Occupational Analysis Division using the Comprehensive Occupational Data Analysis Programs (CODAP). This document is comprised of tables showing the percent time spent, percent members performing, and training emphasis ratings associated with tasks performed in an Air Force career field. In CTO development, the SKT extract is used by SMEs to enhance their perceptions of their career field.

The first step in CTO development is for SMEs to combine STS paragraphs into discrete knowledge areas. After developing the initial knowledge areas, the SMEs then assign item weights to the areas. The weights they assign are based on their evaluation of each area's importance. After the initial weighting, all areas weighted with more than seven items are broken down into smaller, more specific knowledge areas. CTO development involves intense group interaction among the SMEs and the test construction psychologist (TP) and normally takes from 1 to 2 days. During CTO development, SMEs have a large degree of flexibility in determining the CTO format and almost complete autonomy in assigning weights. However, a CTO significantly restricts the degree of flexibility in item writing. Items must test the specific knowledge represented by each knowledge area, and both the TP and the SMEs are charged with ensuring that this requirement is met.

Automated Test Outline (ATO) Development

An ATO begins as a CODAP occupational analysis based on up to 2,000 tasks. The ATO contractor screens the original job inventory and obtains field ratings of testing importance on the remaining tasks. The ATO prepared for the team is comprised of 150 to 200 tasks rated either A, B, C, or D, listed under the "duty areas" used by the occupational survey developer to group similar tasks.

Test developers have minimal involvement in ATO development. The SMEs selected to write the tests participate in the field survey only if they happen to have been in the sample that received the questionnaire from the ATO contractor. After receiving a completed ATO, the test development team must justify any modifications. When an ATO is used in test development, items are referenced to specific ATO tasks. Each item must sample important knowledge needed to perform the specific task referenced. However, the SMEs are the authorities on whether any particular item tests knowledge needed to do a particular task. TPs are not technical experts in the SMEs' career field and must rely upon SMEs to determine how items relate to tasks. This gives SMEs almost total autonomy in determining the content of items developed from ATO tasks.

Method

Subjects

All 10 USAFOMC TPs who had conducted at least one test development project using the current ATO procedures were surveyed. Though this is a small population, it constitutes over half of the test construction psychologists at USAFOMC. All members participated in the study.

Instruments

A survey was developed by reviewing comments on 36 ATO contractor surveys and 17 end-of-project reports submitted by TPs. The survey included 28 statements with a Likert scale for each statement (See Table 1). Space for comments was also provided.

Results

The responses were analyzed with several weighting and averaging methods. However, the clearest method of data presentation was unweighted frequency counts. Therefore, the responses to the survey are shown in Table 1 with this method of presentation. The statements in the table are grouped into seven categories: Task Breadth, Task Differentiation, Task Specificity, Testing Importance Ratings, Group Dynamics, Outline Discreteness, and Overall Considerations.

Discussion

Task Breadth. This refers to coverage of the most important tasks, the critical tasks. Fifty percent of the TPs thought that their ATOs covered the most important tasks in a given career field, and only one TP thought that they did not cover these tasks. In the latter case, the test development team was released from the requirement to use the ATO and subsequently developed a CTO for the project.

The next issue in the category of Task Breadth concerns the number of tasks included on ATOs. Only one TP thought that the ATOs contained too many tasks, and none thought that ATOs contained too few tasks. Therefore, it appears that the number of tasks on ATOs is satisfactory.

The final issues in the category of Task Breadth are more directly related to validity and, therefore, of greater importance. Fifty percent of the TPs thought that the ATOs did not provide complete career-field coverage. These opinions may be interpreted in either of two ways, one positive and one negative. On the positive side, the ATOs in question may have lacked complete career-field coverage because unimportant areas were screened from ATOs. On the negative side, the ATOs really could have been missing important areas.

Table 1. Survey Results

	Strongly Disagree	Disagree	Neither Agree nor Disagree	Agree	Strongly Agree
TASK BREADTH					
ATOs include the most important tasks in a given career field	0	1	4	5	0
ATOs include an appropriate number of tasks for optimal test development	1	0	4	5	0
ATOs include too many tasks for optimal test development	0	6	3	1	0
ATOs include too few tasks for optimal test development	1	6	3	0	0
ATOs provide complete career field coverage	0	5	1	4	0
ATOs include many tasks that are not important in a given career field	0	4	0	4	2
TASK DIFFERENTIATION					
ATO task statements adequately differentiate between E-5 and E-6/7 levels	1	3	3	3	0
TASK SPECIFICITY					
ATO tasks are at an appropriate level of specificity for optimal test development	2	2	2	4	0
ATO tasks are too general for optimal test development	3	3	4	0	0
ATO tasks are too specific for optimal test development	0	2	4	1	3
ATOs are consistent in the degree of specificity throughout duty areas	0	4	3	3	0
TESTING IMPORTANCE RATINGS					
The A, B, C, D method of classifying tasks correctly differentiates tasks of different levels of importance	0	2	2	5	1
ATO task statement ratings (A, B, C, D) adequately differentiate between E-5 and E-6/7 levels	0	2	4	4	0
GROUP DYNAMICS					
Use of ATOs results in more favorable group dynamics with the SMEs than CTOs	3	4	3	0	0
Use of ATOs results in less favorable group dynamics with the SMEs than CTOs	0	0	3	4	3
ATOs and CTOs result in equally favorable group dynamics with the SMEs	2	4	3	1	0
SMEs are more committed to ATOs than to CTOs	4	4	2	0	0
OUTLINE DISCRETENESS					
The ATO's grouping of tasks into duty areas is meaningful for test development	1	1	0	7	1
The ATO's grouping of tasks into duty areas is optimal for test development	2	3	2	2	1
ATOs as received from the contractor are not as discrete as CTOs	0	2	2	4	2
After ATOs are modified by the team they are not as discrete as CTOs	0	2	4	2	2
ATOs allow SMEs more flexibility in item writing than CTOs	0	1	3	5	1
Task organization in ATOs makes identification of overlaps easier than in CTOs	2	0	7	1	0
Use of an ATO results in fewer inadvertent overlaps than use of a CTO	1	2	5	2	0
Production of an inadvertent overlaps is a serious problem when using ATOs	0	5	3	1	1
OVERALL CONSIDERATIONS					
Use of an ATO results in overall project time savings	1	5	1	3	0
Tests developed with ATOs are more job related than tests developed with CTOs	1	3	4	2	0
Overall, ATOs are better than CTOs	2	2	4	2	0

either because the occupational surveys had become outdated, or because the ATO screening algorithm eliminated important tasks.

A related issue is that 60% of the TPs thought that the ATOs included many tasks that were not important in a given career field. It may be that TPs were considering the D-tasks as unimportant. If this were the case, an easy solution would be to simply remove the D-tasks from ATOs. However, eliminating D-tasks without replacing them may solve one problem and create another, having too few tasks. A better solution may be to focus efforts on improving the task screening algorithm and the currency of occupational surveys. This could result in more important tasks being selected to replace D-tasks.

Task Differentiation. Two promotion tests are normally developed during a project, one for promotion to staff sergeant (E-5), and the other for promotion to both technical sergeant (E-6) and master sergeant (E-7). In many career fields, the tasks change with advancement in seniority and responsibility. One issue is the extent to which ATOs distinguish between what is important at the different grade levels. Only 30% of the TPs reported that ATO task statements differentiated adequately between the grade levels. It is possible that inadequate differentiation between grade levels may not be a problem in and of itself. When these opinions are considered together with TP opinions about career-field coverage, this may be interpreted as a symptom of a more basic problem with the quality or coverage of the tasks.

Task Specificity. None of the TPs thought that the ATO task statements were too general, and 40% thought they were too specific. A task that is written in a specific manner may take very exacting and limited knowledge to perform. Likewise, a task that is written in a general manner lends itself to more open interpretation and often requires a greater breadth of knowledge. A related concern was the consistency in the degree of specificity throughout duty areas. Only 30% of the TPs thought that the ATOs were consistent in the degree of specificity throughout duty areas. If an ATO lacks consistency in task specificity, the meaning of the weights assigned to duty areas becomes questionable.

Testing Importance Ratings. These are ratings of task testing importance derived from the group of senior NCOs in the field who complete and return questionnaires to the ATO contractor. Sixty percent of the TPs thought that testing importance ratings correctly differentiated tasks of different levels of importance. Therefore, it appears that the testing importance ratings were adequate to differentiate tasks of different levels of importance. A related concern was the ability of the testing importance ratings to distinguish between the testing importance of the same tasks at different grade levels. The statement that ATO task ratings adequately differentiate between the E-5 and E-6/7 grade levels was agreed to by 40% of the TPs, while 20% disagreed. Given the low number of TPs that disagreed with this statement, it appears that lack of task rating differentiation between grade levels is generally not a problem with ATOs.

Group Dynamics. In this area the results are unequivocal. Seventy percent of the TPs thought that use of the ATO resulted in less favorable group dynamics than use of a CIO. The opposite response pattern was seen to the inverse of the statement. In a similar vein, 80% of the TPs disagreed that SMEs were more committed to ATOs than to CIOs, and none agreed with the statement. These responses suggest that group dynamics and SME commitment can suffer with the use of ATOs. During CIO development, team members not only brainstorm but also get to know each other. This is an intense group interaction which is missing when the ATO is used in test development. Another factor that may affect group

dynamics is the limited ability of TPs to determine how items cover ATO tasks. With an ATO, TPs must ensure each item covers the task referenced. The TPs have limited ability to do this because they do not normally have expertise in performance of the tasks. This tends to make TPs much more dependent on the SMEs. All of these factors may account for the perceived loss of group interaction and SME commitment.

Outline Discreteness. Outline areas are discrete if they do not overlap (i.e., do not allow items to be listed under more than one outline area). While 60% of the TPs found unmodified ATOs to be less discrete than CTOs, only 40% of the TPs found them to still be less discrete after modification. As mentioned above, ATO tasks are grouped by occupational survey duty areas. While only 30% of the TPs found this grouping to be optimal for test development, 80% did find it to be meaningful. These results suggest that there is room for improvement in the discreteness of the areas used to organize ATO tasks. The less discrete an outline is, the more flexibility test developers have in writing items. Sixty percent of the TPs saw ATOs as allowing SMEs more flexibility than CTOs in item writing. One positive result of having flexibility may be production of fewer poor-quality items. Poor quality items may be produced if an outline requires SMEs to write a given number of items from limited reference material. Test development regulations limit reference material used by SMEs to those references readily available to every member of a given career field. If an ATO is flexible, SMEs may choose to write items that are more easily written from their references. The trade-off is that better quality items, from areas well covered in reference material, may be written at the expense of necessary items from important areas that have limited or difficult reference material. However, the value of having better items at the expense of documented and representative career field coverage is questionable. The content validity of promotion tests rests largely upon having representative career-field coverage. Several TPs added comments on their surveys expressing concern that the SMEs using ATOs would tend to write questions which were easier to write. Given TP opinions about ATO discreteness and flexibility, it appears that ATOs may make it harder for TPs to ensure representative coverage of knowledge areas.

Another result of an outline lacking discreteness is the possible production of overlap or duplicate items when writing different sections of a test. Twenty percent of the TPs found this to be a serious problem with ATOs.

Overall Considerations. These are factors which affect the overall utility of ATOs. The first statement listed under Overall Considerations concerns time savings. It was expected that SMEs would spend little time with ATO development, since the ATO is essentially complete when received by the test development team. However, only 30% of the TPs agreed that use of an ATO resulted in overall project time savings. This may be due to ATOs still being in the developmental phase and the occasional need for modifications before use, as well as the fact that this is the first time items have been referenced to tasks.

The next factor is possibly the most important discussed to this point, the job-relatedness of tests developed with ATOs. If most TPs agreed that ATOs improved the job relatedness of promotion tests, this would be an argument that ATOs are an improvement over CTOs. However, only 20% of the TPs thought that the tests developed with ATOs were more job-related than tests developed with CTOs. In any event, a critical facet of content validity is documentation of a clear, direct link between test content and task performance. At first glance, ATOs appear to establish this link between test content and task performance. However, the link is not so clear, as illustrated in the following chart:

ATO task

Repair B1-B radios

Some Related Knowledge Areas

B1-B Radio characteristics
Electromagnetic principles
Security considerations
Soldering

The task as it appears on the ATO is specific. However, items written for this specific task may be based on different types of knowledge related to task performance, since the knowledge to be tested is not specified. Therefore, SMEs may write about any related knowledge as long as they believe a link exists. This approach is the opposite of that taken with a CTO. In a CTO, the knowledge is specified, and the task is unspecified. In both cases, the same weakness exists. That weakness is the lack of a direct link between specific tasks and specific types of knowledges tested. As far as the content validity of the resulting test is concerned, it may be that neither outline is preferable.

The last statement rated was: "Overall, ATOs are better than CTOs." Only 20% of the TPs agreed, while 40% disagreed. These opinions may not be an especially negative reflection on ATOs. TPs may believe they are doing a such a good job with CTOs that ATOs will have to be outstanding to be an improvement. However, the opinions do suggest that ATOs in their current stage of development are not fully accepted by TPs.

It should be noted that the ATOs being considered by the TPs, though similar in format and structure, were unique to the specialties of the SKTs being developed. The results are not as "clean" as they would be if each TP had used the same ATO. For this reason, one must be cautious in interpreting the results. However, this does not detract from the fact that, in some categories, there was strong agreement, agreement which provides strong evidence of real differences of a general nature between ATOs and CTOs.

Recommendations

Most ATO problems appear to be related to outline discreteness, group dynamics, and the quality and coverage of ATO tasks. It may be possible to overcome the first two problems by using more of the SMEs' expertise in ATO development. We recommend using the SMEs to first develop a shell outline of discrete knowledge areas and then to group ATO tasks into these areas. This outline would look like a CTO with the exception that important ATO tasks would be grouped under discrete knowledge areas. This may have several positive effects. First, increased involvement by the SMEs in outline development should enhance their commitment to the outline and may enhance group dynamics. Second, the use of discrete knowledge areas would enable both the SMEs and the TP to more easily ensure representative knowledge area coverage. Again, this may have a positive effect on group dynamics by reducing TP dependence on the SMEs. Perhaps most importantly, documented content validity would be established by the direct link between task performance and the specific knowledge areas tested. Finally, to improve the quality and coverage of ATO tasks, we recommend that future research be directed at maintaining task currency and fine-tuning the ATO task screening algorithms.

The Future of Item Analysis

Howard Wainer¹

Educational Testing Service

Princeton, N.J. 08541

Abstract

This paper reviews the role of the item in test construction and suggests some new methods of item analysis. A look at dynamic, graphical item analysis is provided that utilizes the advantages of modern high speed, highly interactive computing. Several illustrations are provided.

1. Introduction

The individual item has always been the basic building block of a test. As such, a variety of diagnostic tools have been devised to assure that each of these blocks was sound. These tools, taken as a whole, are known as *item analysis*. Historically, the fixed format of tests was a mortar that held the various items firmly together. The predetermined attachment that every item had with its neighbors provided additional strength. If a particular item was weak its adjacent items lent strength that kept the entire edifice of the test from collapsing. An item that was ambiguous when presented individually became less so when it was presented in a particular context. An item that was more seriously flawed had the small recompense of being fair, in the sense that all examinees received the same item in the same context. This uniformity of presentation often allowed subsequent investigation to easily spot the problem, and thus the item and its influence on total scores could then be excised.

Traditional test theory (Gulliksen, 1950) recognizes this molecular role of the item, but emphasizes consideration of the test as a cohesive entity. Thus we might, in modern parlance, rename traditional test theory *Test Response Theory (TRT)*. The second stage in the development of test theory focussed more closely on the item as the fungible unit of test construction. While Lord and Novick (1968) placed TRT into a statistical framework, their inclusion of Birnbaum's work on Item Response Theory (IRT) clearly established this way of thinking of test construction as the wave of the future. Of course IRT did not begin in 1968. Lord's 1952 monograph, *A Theory of Test Scores*, codifying the previous nine years of Lord's work, contained many of the crucial ideas of IRT. Loevinger's (1947) notion of an exam testing a single trait or ability provided the epistemological basis for the most fundamental tenet of the emerging IRT. In 1959 Lazarsfeld furthered the use of latent dimensions with his work on latent structures, and a year later Rasch (1960) published a formal test theory, based on a latent variable, in which the the primary focus was the behavior of the item.

The increasing theoretical weight being borne by the item implied a parallel increase in the importance of items maintaining a uniform high quality. Item analysis grew both in importance and in sophistication. Replacing the useful, but often *ad hoc*, measures of an item's performance (i.e. functions of the proportion of examinees who got the item correct $\{P_i\}$, the biserial correlation between performance on the item and total score $\{r_{bi}\}$, and similar statistics for each of the offered options), were a plethora of analogous measures that related the item to the theoretical construct underlying the test (i.e. the item's characteristic curve and its parameters). Test reliability is currently being supplemented by item and test information functions, which may eventually supplant it. The importance of the plausibility of the strong test model to the accuracy of the inferences made from the scores meant that item analysis now required fit statistics for the model itself.

Emerging awareness of the diverse ethnicity of the test-taking population, and its consequences, added a new load of responsibility onto the item. In addition to the other statistical hurdles that the item must pass, it must also demonstrate that it does not perform differently in a variety of different populations. A number of procedures (Holland and Thayer, 1988; Thissen, Steinberg and Wainer, 1988) have been developed to test differential item functioning (*dif*), yielding still more information that must be obtained and carried in the item analysis.

If current trends continue, in the future we will be leaving behind the fixed format test in favor of computer administered exams that are custom made for the examinee. In such tailored (or adaptive) tests the item analysis will have to inform us of the fungibility of the item. It will have to tell us how dependent the various statistics, which characterize the item's performance, are on the specific context in which the item is presented. A further discussion of these interesting issues is beyond the scope of this account, but the interested reader is referred to Wainer *et al* (in press) for a fuller discussion.

I have recited this litany of responsibilities for the modern item to make explicit the point that item analysis is becoming an increasingly complex enterprise. It must carry the answers to a broad array of questions. Sometimes, the questions we want to ask depend on the answers to previous queries. Also we will often find that the answers of the future will need to be phrased as functions or vectors rather than the scalars of the past. Thus I expect that future item analysis will surely have two characteristics; it will be graphical in presentation and dynamic in character. The rest of this paper will devote itself to speculations and suggestions as to what such a future item analysis system might (or even ought to) look like.

2. Dynamic Item Analysis

It seems clear that, depending upon circumstances, what a test developer wants to see on an item analysis card will vary. It is equally clear that even a small proportion of what one would want to look at will not be easily visible on ordinary size pieces of paper (or an ordinary size computer screen). This is not a new insight. A contemporary test developer has an item card which has the item printed on it. On the back the key is recorded, as is the information on the item analysis strip. Oftentimes there is a paper clip holding one or more sheets of paper with additional information regarding past usage, applicable content specifications, results of previous sensitivity reviews, etc.. Tests are constructed by sorting items into piles reflecting shared characteristics, and then picking and choosing from among these piles to try to satisfy what are often contradictory requirements.

Does this system meld well with the world of modern computing? So far, it hasn't. At least it hasn't with the test developers I know. They use computerized test construction systems as mechanical horses. The computer can dig into its encoded item pool and produce a sample of items for the test developer to choose among. The software usually allows the developer to pick and choose and even has statistical aids like an on-line test information function. Thus if the developer is trying to produce a test that has a prespecified information structure, various item combinations can be tried out. How is this sophisticated system used? Most typically, item types are solicited, and printed out. Then the output is cut up with a scissors and taped on little cards. The item analysis information, which was printed out as well, is also cut up and then is pasted on the back of those same cards. Any ancillary information that the computer might have on the item is then printed out, cut up and clipped to the cards for easy reference. These computer-and-scissor generated cards are then stacked and spread out on a large table. The test developer then begins the task of sorting and test construction.

Surely we can do better!

The reason that cybernetic methods for test construction have not met with overwhelming support is that the interface between program and test developer is often clumsy and unnatural. Our quest for better item analysis methods and presentation runs smack into these same issues. It seems to me that if the card metaphor is natural and easy for test developers, we ought to follow it in developing its computerized realization. The item should appear on an electronic card that has a variety of "buttons" on it. The user "presses" the button of interest and immediately the material promised becomes visible. Thus a reader might push the "Option trace line" button to view the performance of the item's choices in the context of the entire test. Another button might yield a look at the traditional statistical display; still another could give the key. This is a natural way for a developer to explore an item's characteristics. One such electronic item card is shown in exhibit 1. This is taken from an experimental system based on the commercially available computer program HyperCard (Atkinson, 1987; currently only available with the Apple Macintosh) that I have invented for the purposes of this article. It may have some additional uses soon.

Insert Exhibit 1 about here (Hypercard 1A)

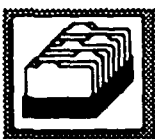
An explanation of the HyperCard Item Analysis System

The main part of the screen is devoted to portraying the item. The arrows beneath the item will move the viewer to the previous item (pointing left) or the next item (pointing right). The curved arrow returns the viewer to the beginning.

The "Notes" section contains any information you like: text, graphics or both. In this instance it alerts the viewer to the fact that this item was flawed, the nature of the flaw, and a graphical notation providing an extra reminder.

The "item statistics" section contains the information indicated, providing for both traditional and IRT interpretations. The sample size from which these statistics were calculated indicates clearly that these are accurately determined.

The ten small icons at the bottom are electronic "buttons". These are activated by the mouse and each have a different function (which I elaborate on shortly). The notion is that each of these provides something special which may (or may not) be of use at any given time. Activating these buttons brings this additional information to the screen.



Brings the viewer to the test file box (more about this later). In this box are kept the items which make up the total item pool, organized by subject area, or anything else a particular test assembler might require




This is the item "key." It contains the keyed correct response as well as the explanation as to why this is so. For mathematics items this might include a derivation or proof; for verbal items a detailed discussion of each option. It also contains information about the item's history; its author, its reviewers, etc..



Activating this button brings to the screen the item's option trace lines — plots of $P(\text{option} | \theta)$ vs. θ . Or in traditional terminology, the "option raw score regressions" — plots of $P(\text{option} | \text{raw score})$ vs. raw score.



This button is much like  except that it provides plots conditioned in the other direction (as we discussed previously). Thus these are $P(\text{score} | \text{option})$ vs. score; or in an IRT framework, $P(\theta | \text{option})$ vs. θ .



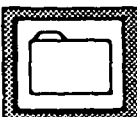
This button provides the user with a test scores by options table, in which the entries are the number of examinees with a given score who chose each option. Obviously, it is from a table like this (in which scores can be raw scores, equated scores, or IRT estimates of proficiency) that the figures described above are constructed. There is a button on this card which shows the fifths table.



Pressing this button writes all of this test information into a single file that can easily be copied onto other sorts of electronic media. It can also be used to transfer item information into other programs for further statistical analyses not included in this package.



This button brings up a screen containing the item's content characteristics that would be used to satisfy the test specs.



This button is the active test folder. It contains all of the items that are actively being considered for inclusion in the current test. When pressed it shows a folder with all the items so far selected, along with a variety of test statistics (total test information, extent to which test specifications have been fulfilled, etc.).



This button indicates the intention to measure the extent of differential item functioning (*dif*) that this item possesses. Pressing this button brings up another screen which inquires about the make-up of the groups for which this item's performance is to be compared.



This is a special purpose button useful for certain kinds of music items ("name that tune" items). When activated it will play the item. It is included here to illustrate special purpose buttons that can be included to aid the test construction process. Alternative buttons might just as easily be used to read passages in tests of aural comprehension or spelling. I expect that a similar button ought to be included to bring to the screen items (like those testing reading comprehension) that are too large to fit in the allotted space on the item card. Modern computer screens (such as the one I am using as I write this paper) can easily legibly contain two full pages of text. It would be a small matter to allow the test developer to bring up longer items onto such a screen with a mere flick of the mouse (or other response device).

Putting the items together

The active test folder described in the previous section is the next step in test construction. It is the place in which the items, once examined and qualified, are accumulated into a test. Whether this is a test that will be administered "as constructed," or an item pool to be selected from by some sort of adaptive algorithm, is immaterial at this point. One can imagine that the test folder (one such is pictured in exhibit 2) is divided by subclassifications of subject matter, and as items are placed in it a variety of "whole test" indicators (i.e. the amount of test information and the extent to which test specifications have been satisfied) are continuously updated. While this is a fascinating area of concern, it is sufficiently beyond my charge in this paper that I will leave a fuller description to other accounts.

Insert Exhibit 2 about here (Hypercard test file)

3. Discussion

So far this article has been reminiscent of Stephen Leacock's young man who ran out of his house, jumped on his horse, and galloped off in all directions at once. One reason for this appearance has been that item analysis is a tool that must serve many masters; a second is that many users of this tool are a long way from being masters of it. Some of the components of item analysis are, by their very nature, statistically complex. Test developers must be, first and foremost, experts in the subject matter of the exam. Formal psychometric training, if it is there at all, comes later. This is as it must be I'm afraid. I have seen the alternative and it was not pretty. I know of at least one test (happily, not an ETS product), made up by trained psychometricians, whose statistical characteristics were impeccable, but whose content was problematic at best; see Wainer and Kiely (1987) for a discussion of some of these problems.

Why graphs?

Thus a usable item analysis must take into account the inevitable fact that some of its users will be taking the numbers merely as heuristic indicators — rules of thumb that aid in obtaining an acceptable test ⁽¹⁾ — without any deep understanding of the technical aspects of their underlying meaning. One purpose of this article is to emphasize that graphical heuristics are often easier to understand and more difficult to misuse because they can carry their own context with them. This easy? Not quite, for graphical summaries are a bit more difficult to concatenate. The ease of characterizing a test by its average r_{bis} is lost when one "merely" has a test full of curves. But is such a single numerical summarization proper? Einstein reminded us that "everything should be as simple as possible, but no simpler." Modern test theory has taught us that the basic analytic characteristic of an item is a function, not a scalar. Items cannot be statistically characterized in isolation. To know an item's difficulty we must know who is its audience, what kinds of items are presented with it, what came before it and a myriad of other details. Thus it is both proper and useful to think of item characterization in functional terms. But functional analysis is a relatively esoteric subject; few readers of this paper are likely to be familiar with the kinds of p-dimensional mappings in Hilbert space that are the hallmark of discussions in this domain⁽²⁾. But the representation of functions in a graphical form allows even the mathematically unsophisticated access to the power achievable with this approach.

Graphical representation can be used more broadly than I have so far sketched. This is only a beginning. One of the great unsolved problems in test development is how to characterize an item's content more compactly than merely reproducing the item. For example, we could characterize the item:

$$3 + 6 = ?$$

as: "arithmetic, integer, addition." But what about a verbal item? Perhaps a graphical metaphor (a Chernoff face?) could be used to characterize content? Recent modifications to these meta-iconic displays (Flury and Riedwyl, 1981) allow the compact display of more than 20 variables. Using such a method might allow the test developer to get a reasonable understanding of the content of a test with relative ease. It would also further blur the distinction between content validity and face validity.

Providing plots of an item's content is a tough problem. Tougher still is the problem of measuring the extent of an item's fungibility. How much does an item's essential character change with context? We could make a plot of an item's difficulty against variations in context, if we knew how to measure context. A start on this was provided by Dorans and Kingston's (1985) analysis of the effects of item location. Should we study the robustness of item statistics across contexts, e.g., across permutations of item ordering? I welcome all suggestions on this one.

Why Item Analysis?

Item analysis is just one tool, albeit an important one, in the workshop of the test developer. Hence in this article I have often strayed from narrow discussions of item analysis to the broader topic of test construction. It is important that we keep in the front of our minds that the construction of high quality tests is the *raison d'être* of item analysis. In the development and assessment of new item analysis techniques we must ask three questions:

- 1) What information does the test developer need?
- 2) How can we best characterize this information?
- 3) How can it best be communicated?

I believe that the answer to each of these requires that we stress graphics and flexibility. A test developer may have many potential questions about an item in mind as the process of test construction proceeds. Exactly what sorts of information required will vary from item to item, and from situation to situation. Everything must be easily available, but we must not clutter things up with the unnecessary. We rarely suffer from an information overload; more frequently it is a non-information overload. Thus I have structured the HyperCard item analysis in such a way so as to provide huge amounts of information a mere button push away. Yet the developer need not sift through them, if they are not required.

Last, the most wonderful and complete item analysis methods will not be used if they are not accessible to test developers. We must not decry the limitations of psychometric training found among most test developers. Just as they ought not to blame us for not knowing the intricacies of their areas of expertise. Instead we must recognize those limitations and build computer systems⁽³⁾ meant for this class of users. Again, my HyperCard system is a first step toward such a system. This step was aided enormously through my contact with ETS test developers. I hope the ideas expressed here lead to something that partially pays them back for their help.

4. Summary

Explicitly restated, here are the main strands of my thoughts in this area:

(1) Item statistics conveyed in an item analysis are crucially related to context. As such it is natural to think of these summaries in functional form. Functions are best represented graphically and so an inescapable conclusion of this syllogism is that **graphical summaries must play an important role in any future item analysis.**

(2) Item analysis must serve many purposes, but foremost is as a handmaiden to test developers in the building of tests. Thus any **item analysis techniques must meld gracefully with the tasks and practices of test development.**

(3) **Flexibility is crucial in item analysis.** To effectively display items in some tests we might want to play music, others (spelling, foreign language dictation, etc.) could require spoken language, still others need complex drawings. Any item analysis scheme must leave room for easy customization, taking advantage of the capability of modern computing. I included only one "off-the-wall" button in exhibit 6: the music button. But many other special purpose buttons are conceivable. The mind set of flexibility is important.

(4) **An item analysis/test development system must be easy to use.** Even here at ETS, a hotbed of test construction, no one makes up a test more often than once every six weeks. A lot can be forgotten in six weeks. Any usable system must be sufficiently intuitive so that an expert in music or English literature can sit down at it and use it after a substantial hiatus. Anything less will provoke anger⁽⁴⁾ and so will ultimately fail.

References

- Atkinson, B. (1987) *HyperCard*. Cupertino California: Apple Computer.
- Dorans, N. and Kingston, N. (1985) The effects of violations of unidimensionality on the estimation of item and ability parameters and on item response theory equating of the GRE Verbal scale. *Journal of Educational Measurement*, 22, 249-262.
- Flury, B. and Riedwyl, H. (1981) Graphical representation of multivariate data by means of asymmetrical faces. *Journal of the American Statistical Association*, 76, 757-765.
- Gulliksen, H. O. (1950) *A Theory of Mental Tests*. New York: John Wiley & Sons. (Republished in 1987 by Lawrence Erlbaum Associates of Hillsdale, N.J.).
- Holland, P. W. and Thayer, D. (1988) Differential item performance and the Mantel-Haenszel procedure. Chapter 9 (pps. 129-145) in H. Wainer & H. Braun, (Eds.), *Test Validity*. Hillsdale, N.J.: Lawrence Erlbaum Associates
- Lazarsfeld, P. F. (1959) Latent structure analysis. In S. Koch (Ed.) *Psychology: The study of a science*, Vol. 3. New York: McGraw-Hill.
- Loevinger, J. (1947) A systematic approach to the construction and evaluation of tests of ability, *Psychological Monographs*, 61, No. 4.

- Lord, F. M. (1952) A theory of test scores. *Psychometric Monographs No. 7*, Psychometric Society.
- Lord, F. M. and Novick, M. R. (1968) *Statistical Theories of Mental Test Scores*. Reading, Massachusetts: Addison-Wesley.
- Ramsay, J. O. (1982) When data are functions. *Psychometrika*, **47**, 379-396.
- Rasch, G. (1960) *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Nielsen and Lydiche.
- Thissen, D., Steinberg, L. and Wainer, H. (1988) Use of item response theory in the study of group differences in trace lines. Chapter 10 (pps. 147-169) in H. Wainer & H. Braun, (Eds.), *Test Validity*. Hillsdale, N.J.: Lawrence Erlbaum Associates.
- Wainer, H. (1983a) Pyramid Power: Searching for an error in test scoring with 830,000 helpers, *The American Statistician*, **37**, 87-91.
- Wainer, H. (1983b) Reply to Oderwald and others. *The American Statistician*, **37**, 351-352.
- Wainer, H. and Kiely, G. (1987) Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, **24**, 185-202.
- Wainer, H. and Braun, H. (1988) *Test Validity*. Hillsdale, N.J.: Lawrence Erlbaum Associates

Footnotes

[§] This is a draft of an invited article that will be appearing in the *Journal of Educational Measurement*, all references should be made to that source. This work was supported in part by the Educational Testing Service; I am grateful for that as well as for the support and encouragement of Henry Braun, C. Victor Bunderson, Paul Holland, Charles Lewis, Mari Pearlman, Nancy Petersen, David Thissen, Wendy Yen and many members of the ETS test development staff.

⁽¹⁾ And some, particularly litigators, will be using item analysis numbers out of context; selecting the pieces of the analysis that serve their narrow purposes and ignoring other aspects.

⁽²⁾ An eloquent discussion of the use of, and need for, functional analysis in social science applications is contained in Jim Ramsay's (1982) presidential address to the Psychometric Society.

⁽³⁾ I refrain from utilizing the hackneyed phrase "expert system" here, although that is what we should aim for. I believe that we are still far away from knowing enough to build such a system in most subject matter areas: maybe there is a chance in arithmetic.

⁽⁴⁾ My colleague, Mari Pearlman, added here "and frustration and time loss and inefficiency and inaccuracy — which are all considerably more important than anger."

Exhibits

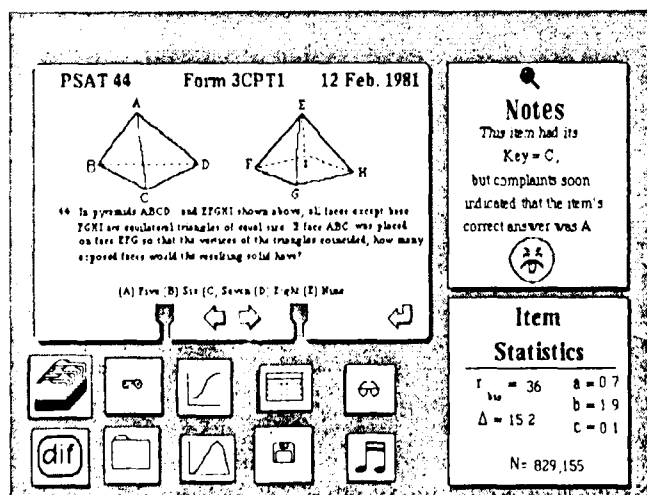


Exhibit 1

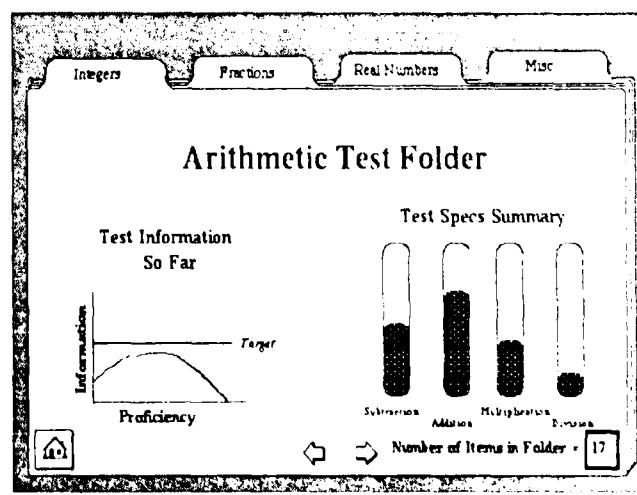


Exhibit 2

The Buros Institute of Mental Measurements in the 1990s

Barbara S. Plake, Director
Buros Institute of Mental Measurements
University of Nebraska-Lincoln

History and Tradition of the Buros Institute of Mental Measurements

More than 50 years ago, Oscar Krisen Buros recognized the need for an organization that would "test the tests" and provide consumers with current, critical evaluations of test materials. Towards this end, Oscar Buros published the first volume of test reviews in 1938. Over the next 50 years, a total of nine Mental Measurements Yearbooks have been published, containing reviews of thousands of commercially published tests.

The Buros Institute of Mental Measurements was moved to Lincoln, Nebraska in 1979 and it exists today as an integral part of the Department of Educational Psychology of Teachers College of the University of Nebraska-Lincoln. The objectives of the Buros Institute remain almost identical to those established by Oscar Buros half a century ago. These objectives include the publication of descriptive information and candidly critical, scholarly reviews of tests and test-related products published in the English-speaking countries of the world as well as advancement of the applied science of educational and psychological measurement. As tests become more technically complex, and as the proliferation of commercially available tests continues, the need for dissemination of critical evaluative information about the utility of commercially available tests becomes even more necessary.

Current Products of the Buros Institute of Mental Measurements

Mental Measurements Yearbook and Tests in Print Series. The major undertaking of the Buros Institute is the publication of the Mental Measurements Yearbook (MMY) and the Tests in Print (TIP) series. The MMY series contains critical test reviews written by qualified professional people representing a variety of viewpoints. Tests in Print is an indispensable part of the MMY series as it represents a comprehensive bibliography of all known commercially available English-language tests. It serves as an index to all in-print tests that have been reviewed in previous MMYs.

Mental Measurements Yearbook Database. In order to promote enhanced communication about new test products, the Buros Institute contracted in 1983 with Bibliographic Retrieval Service (BRS) to prepare and support a searchable database of test reviews. Users of the Mental Measurements Yearbook Database (MMYD) on BRS will experience two outcomes: (1) early access to test reviews prior to their inclusion in the MMY, and (2) the ability to combine search algorithms to achieve efficient information on instrument availability and quality. For example, individuals accessing MMYD are able to retrieve information such as test name, test classification (e.g., achievement batteries, language, neuropsychological, reading, intelligence, vocational, etc. based on Mental Measurements Yearbook classification schemes), test author, test publisher, publication dates, population served, scores indices, administration information, reliability and validity data, and price. By using the search environment, questions such as the following could be answered:

- * What individually administered speech and hearing tests are appropriate for six- to eight-year-old children?
- * Has Science Research Associates published any tests which might be appropriate for the selection of school personnel?
- * Which behavior checklists/inventories have been reviewed in reviewed in the MMY series?

Therefore, through accessing MMYD on BRS individuals can obtain not only professional reviews but also can answer important questions about the availability and appropriateness of various test instruments.

MMYD contains all reviews prepared by the Buros Institute starting with the Eighth MMY and is updated monthly. Specific information regarding access to the MMYD is available from BRS Information Technologies, 1200 Rt. 7, Latham, NY 12110 (800-458-0908).

Supplement to the Mental Measurements Yearbook Series. Although the database provides consumers with an excellent resource of evaluative information about tests, still a large number of test users are not able to take advantage of this improved information source due to their lack of availability to appropriate technology. Therefore, the Buros Institute adopted a more ambitious publication schedule and product line. In 1988, The Supplement to the Ninth Mental Measurements Yearbook, the first in the Supplement series, was published. Volumes in the Supplement series will contain printed versions of test reviews prepared since the publication of the most recent Mental Measurements Yearbook.

New Publication Schedule. Beginning in 1989, a hard-bound MMY will be published every other year. These volumes will be edited by Jane Close Conoley and Jack J. Kramer. In years in between publication of a MMY, a soft-bound Supplement will appear. Consumers with critical needs for current test reviews who do not have access to the computer database will find the Supplement a valuable information source. Other users (for example, those who have ready access to MMYD) may be able to postpone access to current test reviews until they appear in the hard-bound volume that will appear in the following year. All reviews published in the Supplement will also be printed in the subsequent MMY, along with approximately 70% new reviews.

Current Programs of the Buros Institute of Mental Measurements

Buros-Nebraska Symposium on Measurement and Testing. Each year the Buros Institute sponsors a symposium dealing with a critical issue in the measurement and testing fields. A list of previous symposium topics follows:

1. Social and Technical Issues in Testing
2. The Future of Testing
3. The Influence of Cognitive Psychology on Testing
4. The Computer as Adjunct to the Decision Making Process
5. Assessment of Teaching: Purposes, Practice, and Implications for the Profession
6. Curriculum-Based Assessment: Examining Old Problems, Evaluating New Solutions

Subsequent to each symposium, a volume containing manuscript versions of presented papers and additional solicited chapters is published through Lawrence Erlbaum and Associates.

Applied Measurement in Education (AME). AME is a scholarly journal created to provide both a greater understanding of educational measurement issues and an improved use of measurement techniques in education. Its intended audience consists of both researchers and practitioners who are interested in research that has a likely impact on educational measurement practice. Sponsored by the Buros Institute, AME is published by Lawrence Erlbaum and Associates.

New Projects and Products from the Buros Institute of Mental Measurements

Reviews of Computer-Based Test Interpretations. As the field of measurement and testing is expanding to incorporate computer technology, a new field of test development has emerged which provides computerized interpretations of psychometric instruments. Identified as being substantially different in scope from the commercially available tests reviewed in the Mental Measurements Yearbook series, a separately edited volume dedicated to evaluative reviews of computer-based test interpretive instruments is being prepared. Edited by Collie Conoley, this volume is scheduled to be published in 1991.

Buros Memorial Library. By virtue of reviewing commercially available English-language tests, the Buros Institute has a comprehensive collection of test instruments and test-related products. These tests, along with Oscar Buros' historical collection of measurement texts and products, is housed on site at the Buros Institute of Mental Measurements. Also contained in the library are measurement-related publications and journals. One goal of the Institute is to increase the scope and accessibility of the Buros Memorial Library. A featured initiative to seek collections for the library is forthcoming.

Professional Consultation. Through the Buros Library and professional consultation, the Institute provides information to governmental agencies, public schools, and individuals. Although not in the business of recommending tests, the staff does attempt to make the resources of the Institute available to those who want to learn about specific tests and testing practices.

Summary

The purpose of this paper was to provide the Military Testing Association information about the Buros Institute of Mental Measurements. The primary product of the Buros Institute is the Mental Measurements Yearbook series containing candidly critical test reviews written by professionals in the field of education and psychological measurement. One upcoming dimension is a more frequent publication schedule for the MMY. The Institute sponsors an annual symposium dealing with a critical issue in the measurement and testing fields. Through the Buros Library and professional consultation the Institute also provides information to governmental agencies, public schools, and individuals. Further, in conjunction with Lawrence Erlbaum and Associates, the Buros Institute sponsors a scholarly journal dedicated to the application of educational and psychological measurement research to the educational process, Applied Measurement in Education. The Mental Measurements Yearbook Database, offered through BRS Information Technologies, provides searchable access to completed reviews even before they are published in the Yearbook.

Changes in the Institute are aimed at broadening the scope of the measurement activities and at improving the current service to the measurement community. The Buros Institute is expanding its measurement-related activities in ways that aid in the better use of tests and testing practices. By providing professional assistance, expertise, and information to consumers of commercially published tests, the Institute hopes to foster meaningful and appropriate test selection, use and practice. Additionally, the Institute hopes to encourage improved test development and measurement research through thoughtful, critical analysis of measurement instruments and the promotion of an open dialogue regarding contemporary measurement issues.

ITEM BANKING IN SKILL QUALIFICATION TEST (SQT) DEVELOPMENT

ALLAN L. PETTIE

U.S. ARMY TRAINING SUPPORT CENTER

FORT EUSTIS, VA

PURPOSE

This paper described the Individual Training Evaluation Directorate's (ITED) efforts to automate, through item banking, SQT development, revision, and production. Following an extensive review of the Individual Training Evaluation Program (ITEP) during 1986-7, TRADOC concluded that ITEP should be allowed to stabilize and made numerous recommendations for improvements. One recommendation was the systematic accumulation through computerized means of test items whose worth has been established through use. ITED examined commercially available software, selected Test Generation System (TGS) marketed by Tutorial Systems, Lexington, Kentucky, and is pilot testing the software.

BACKGROUND

SQTs, begun in 1977, are used by the U.S. Army to test enlisted soldiers in their ability to perform selected tasks of Military Occupation Speciality (MOS). SQTs can be considered to be minimal competency examinations and are designed to serve two major purposes: (1) to provide scores for use in personnel decisions, including the selection of soldiers for promotion to the next higher rank and for retention in the Army, and (2) to provide feedback useful to commanders in training soldiers. The SQT has become almost entirely a paper-and-pencil multiple choice test.

SQTs comprise one component of the ITEP. The ITEP evaluates soldiers' proficiency in Soldier's Manual tasks by three methods. The Commander's Evaluation is a hands-on assessment on tasks related to unit mission. The Common Task Test is a hands-on test of basic survival and combat tasks. SQTs cover MOS skill level tasks. Approximately 800 SQTs are developed, primarily through manual procedures, each fiscal year. Development of these tests is the primary responsibility of 21 U.S. Army proponent schools (training centers) with assistance from the U.S. Army Training Support Center (ATSC), Fort Eustis, Virginia. With the onset of automation, ATSC commissioned a

The views, opinions, and/or findings contained in this report are those of the author and should not be construed as an official Department of the Army position, policy, or decision.

study under the Army's Scientific Services Program of SQT development to determine how SQT development can be improved. This study, conducted by Dr. Anthony Nitko of the University of Pittsburg, made numerous recommendations to improve SQT development and many of these recommendations pertained to the use of item banking and automation. One recommendation was to select a commercially available item banking package and pilot test it.

ITED examined commercially available software and selected the TGS software package which best suited SQT development needs and also would not require additional hardware procurements. ITED procured this software, which operates on IBM compatible microcomputers, which are becoming increasingly available in the proponent schools. Modifications were made to the software to customize it to the existing manual system and to include familiar terms.

SQT DEVELOPMENT NEEDS

SQTs are developed following procedures set forth in TRADOC Regulation 351-2, Skill Qualification Test (SQT) and Common Task Test (CTT) Development Policy and Procedures. These procedures include guidance for task selection, item development, soldier tryout, camera ready mechanical production, and numerous other required SQT development elements. The 21 proponent schools follow these standardized procedures in a variety of ways. Some develop SQTs completely manually, some obtain word processing support or maintain items on word processors, some have data processing support for soldier tryouts, most all have graphics support for graphics and camera ready mechanicals, some lack secretarial support, and some are almost entirely automated. Because of the varied types of printers, microcomputers, scanners, and other hardware available to SQT development branches, the software must be capable of supporting numerous kinds of hardware.

Perhaps unique to SQT development is the requirement to tryout trial SQT task tests and items on performers and nonperformers in the field to provide task standards, identify flawed items, set time limits, and set the minimum passing score for the SQT. These procedures require the manual calculation of proportion correct by performer and nonperformer groups, task score for each performer and nonperformer, time for each performer, and numerous others. Reduction of the calculations is one goal of the item banking project.

SQT task tests are developed by the proponent for each particular task. Many tasks are performed by MOS holders from other proponent schools (i.e., an Infantry soldier may be responsible to perform an Armor School task). The nonproponent school must request these tasks from the proponent if the tasks are selected for inclusion in an SQT. Thus, there is a definite

need to export and import items with complete documentation between schools. This procedure is a completely manual process currently. Also, those schools which have word processing files need not duplicate their effort. Batch loading of these items, current SQTs, statistical data, and other information is another goal of the item banking system.

TGS PILOT TEST

ITED selected TGS for pilot testing at four proponent schools - Aviation Logistics, Fort Eustis, Virginia; Chemical, Fort McClellan, Alabama; Infantry, Fort Benning, Georgia; and Intelligence, Fort Huachuca, Arizona. Initial feedback from the pilot test schools indicated that some terms (e.g., competency, classification) were confusing, there was a need to process soldier tryouts, produce ATSC required forms (e.g., Template Summary Report), transfer items and data among schools, transfer data from ATSC to schools, produce SQT specific item analysis and related statistical reports, and process field inquiries efficiently. The goal of the pilot test is to develop a totally integrated system which will provide, but will not be limited to, the following features: (1) Complete word processing capability. (2) Unlimited field size for storage of items, references, comments, task and item situations, test aid data, graphic aids, and other data elements. (3) Production of draft and final copy tests through a variety of printers (i.e., dot-matrix to off-set inclusive). (4) Efficient scanning of graphics, text, and answer sheets. (5) Production of answer key (template) and other ATSC required reports. (6) Production of SQT specific statistical reports. (7) Capability to export and item task tests, items, and supporting data between schools. (8) Capability to denote items under revision, not to be included with other items on the same test, to be deleted, and lacking soldier tryout data. (9) Capability to classify item by task, by subject area, situation, SQT, MOS, and type (multiple-choice, matching, etc). (10) Capability to select items by statistics, type, task, and subject area or at random. (11) Capability to select tasks by subject area, SQT, MOS, and situation. (12) Capability to present graphics aids, situations, and tables with tasks and/or item. (13) Capability to record references and justification for correct answer with item. (14) Capability to record subject matter experts comments about items. (15) Complete statistical information stored with item. (16) Complete computer based training and user's manual. (17) Security system to prevent item and test compromise.

The TGS pilot test is scheduled to be completed April 1989. Based upon the system developed up to this time in the pilot test, few software modifications are needed to achieve the above stated features. The major deficiency identified thus far appears to be a lack of computer literacy on the part of SQT developers. SQT developers will have to be trained not only on item banking/test generation software but also operating

computers. The transfer of data from ATSC to the schools will required computer skills and training. Thus, the major outcomes of the TGS pilot test have been the development of an integrated, automated SQT development system to duplicate the manual system, the identification of training needs, and the feedback necessary to modify full implementation plans.

Training Test Item Developers: A New Approach

Harvey Rosenbaum
Communications Technology Applications, Inc.

Introduction

Traditionally, people are taught how to develop language test items by means of a general "immersion" method. Novice item writers are presented the desired item format(s), provided some training in potential problems with items, and given a lot of loosely structured, hands-on practice developing items. Many "sink", but some "swim" and become good item writers. While unquestionably demonstrated to work, this approach is neither efficient nor does it optimize the potential in peoples' skills and abilities.

The usual rationale for the "immersion" method is that item writing is a sophisticated task, drawing on higher-order processes and requires a complex interplay of knowledge, skills, perceptions, techniques, and judgments which can not be easily or directly imparted. A perception that echoes the "good teachers are born and not trained" view of life.

This paper describes a course for training item developers which pursued a different approach. The course was developed and piloted for the Defense Language Institute Foreign Language Center (DLIFLC) as part of their program for developing the Defense Language Proficiency Test IV (DLPT IV). It accepts as fact that item writing, like any complex task, will not be completely captured or reflected in a linear sequence of procedures or training steps. But it also assumes that the majority of skills, knowledge, and techniques involved can be identified, introduced, exemplified, and practiced both as discrete operations and as part of sub-task complexes.

This approach also assumes that a knowledge framework that enables students to recognize and relate what they are being trained-in provides a significant advantage to the learner. The use of controlled learning in the context of students understanding the "why" and "what" of instruction is certainly not new; its application to complex, higher order tasks is.

This paper presents a brief description of the training course, the approach used to develop the course, the methods used for instruction and training, and the preliminary results based on the pilot course.

Description of the course

The goal of DLPT IV is to deliver a set of multiple-choice, computer based foreign language tests to assess the language competency of DLI students. The goal of this course is to train potential item developers in the most current knowledge, techniques, and methods for developing language test items for the DLPT IV.

The main skill areas emphasized in the course are developing good stems and options, text selection and editing, and the mechanics and procedures to be followed at DLI in developing items. These practical elements are learned

in the context of the principles and concepts that reflect the cognitive processes underlying the activity of item development.

The two weeks of course instruction are divided into 12 units:

- Introduction to course and DLPT IV
- Item Format
- Writing different kinds of items
- Mastering the mechanics
- Principles of item writing
- Developing items
- Listening comprehension items
- Reviewing items
- Selecting texts
- Editing texts
- Assembling the model test
- Wrapping-up

Course materials consist of a detailed Instructor Guide, Participant Manual, more than 100 worksheets and handouts (in both paper and vugraph form), and tapped oral items. The majority of worksheets and handouts are test items carefully designed to represent specific points and illustrate specific problems which are summarized by 30 principles for writing good test items. These principles cover all aspects of the item including the passage, stem, distractors, key etc.

The participant, or trainee population, for the pilot course consisted of six teams of DLI foreign language instructors. Course materials and examples were developed in English. Trainees developed some test items in English, but primarily in their native language in order to bridge to their target language. They developed items that test the comprehension of factual, inferential, and cultural information at various levels of difficulty.

The course design calls for the introduction of new information and material in the morning and the hands-on application and practice of the morning's learning during the afternoon. However, due to administrative reasons we were unable to implement the afternoon phase of this plan. DLI trainers observed and participated in the morning instruction and met with the course developer in the afternoon to provide feedback and preview the following class.

Framework for developing the course

The orientation that guided the development and structuring of the course can be summed-up by the question "What kinds of awareness, understanding, and knowledge does a good item writer possess?" This question can be answered many ways, but we chose to frame the answer from the point of view of the examinee and therefor assumed that selecting an option to a test question is fundamentally a cognitive process. We also assumed that good item writers have tacit knowledge of many of the elements and components involved in this process.

It then follows that good item writers possess an understanding of how:

- * examinees go about obtaining information from the printed page or speech input
- * examinees construct necessary and specific meanings
- * examinees are affected by the absence or presence of specific information in their selection of options

Part of this information, particularly the more narrow skills and observations, is supplied by the 30 specific principles for writing good items. Most of these principles address commonly recognized problems that novice item writers frequently make.

But part of the answer also comes from current research in psycholinguistics, linguistics, and information processing. The keystone principle in this course is that comprehension is a constructive process. Simply put, the comprehension of written and oral information is accomplished through the processing of input (print or sounds) and its integration with the relevant and necessary stored information in the examinees head. Course trainees dramatically experience this process by reading one of the classic Bransford and Johnson (1972) type passages. The example used is:

The procedure is quite simple. First you arrange things into different groups. Of course, one pile may be sufficient depending on how much there is to do. Otherwise, you are pretty well set. It is important not to overdo things.

Following extensive discussion of how much they can recall of the passage and their experiences in attempting to understand this passage, they reread the passage with the crucial knowledge of the title -- Washing Clothes.

The "comprehension is a constructive process" principle also establishes the framework for teaching students to identify and distinguish between factual, inferential, and cultural information based on the source and type of information being processed. The point is made that all of these forms of language comprehension require a constructive process. What distinguishes them are the differences in how much information and what types of information must the examinee draw upon from internal storage in order to select the correct answer.

For example, factual test items require only that the examinee be able to correctly interpret the presented information in order to select the answer. In this process, the correct interpretation of the input information enables the examinee to directly construct a passage meaning which contains all of the information for the selection of the best option.

Inferential items do not contain all of the necessary information and therefore require that examinees supply crucial information from their own heads. An example was the Washing Clothes passage. In this case an impossible task because the passage did not contain any of the kinds of information that could be used to locate in storage what was missing.

Cultural items are usually also inferential, but require that the examinee possess information that is very specific to a culture, situation, or period. For example, an unforgettable bumper sticker that appeared in the late 70's:

Will Rogers never meet the Ayatollah.

In developing tests exclusively for native speakers, the cultural issue takes on a slightly different cast and becomes one of education level, socio-economic background, or simply age.

Instructional methods

While much of the content and even the concept of this course is new, much is based on traditional and proven methods of item development and training. The essential elements of item writing are introduced and practiced in a planned sequence, combining practice in more than one element when feasible and always helping the student to keep in mind the end goal -- quality test items. The critical mass in skill development is achieved in the units dealing with Principles of Item Writing and Developing Items.

Other methods include:

- * repeated presentation of specific examples of good and faulty items so that trainees experience these items as part of a process
- * interactive sessions between instructor and participants and between participants
- * demonstrations and work problems for hands-on experience and learning

A repeated theme in the course is that good item writers understand what it is like to be an examinee responding to their items. Throughout the course, trainees are involved in exercises and activities that put them in the shoes of the examinee. This experience, combined with the theoretical orientation, makes trainees more aware of what they are requiring of the examinee as they develop a test item. The effect is to give trainees better and more accurate control of their item writing skills.

Trainees, as developing item writers, are also seen as a valuable resource that can be used to help one another learn to think as item writers. While the course itself is intended to achieve this goal, trainees contribute toward this goal through techniques designed to help them tell in detail what they were thinking when they made specific decisions in developing an item. The focus is on explaining decisions, the information surrounding the decision, and rationales.

Results

The pilot version of this course was extremely well received by the participating language teachers, observers, and DLI staff. Unfortunately, the assessment of results can only be based on this experience since the

course has not yet been repeated; the next presentation is planned for December. Participants feedback, both informally and through the end-of-course evaluation questionnaire was typically very positive. Participants found the theoretical framework to be helpful and stimulating. As one participant wrote, "the actual application of the theories made it easier for me to remember the discussions."

The course was rated superior in holding their interest and was unanimously recommended for other language teachers. Among their many accomplishments, participants noted that they learned to how to develop more effective test items, how to avoid pitfalls, how to select material, and how to determine what an item is testing.

The major suggestion on improving the course dealt with providing more time to develop items and receive feedback. This quite legitimate concern resulted from the fact that the afternoon practice workshop had to be cancelled because of teaching responsibilities. Consequently, the item development practice was constrained to the morning session and homework assignments.

An unanticipated, motivating element also contributed to the success of the course. The trainees, or language teachers, had not received formal training in developing language tests and test item; but as part of their teaching responsibility developed weekly tests to assess the progress of their students. Here was an immediate and very practical application for what they were learning in the course.

Preliminary results suggest that this course provides a more efficient and effective method for training items writers. It is also likely that the approach can be modified and applied to the development of training courses for other kinds of complex, cognitive tasks. The basic elements of the approach are:

- * the cognitive processes required to conduct the task are the point of departure
- * develop a cognitive model to provide a framework and road map through complex terrain
- * identify the more concrete / discrete elements trainees must master
- * use concrete examples to focus the trainees on the main points
- * create examples that provide trainees with experiences that allow them to explore and become aware of the distinctions, insights, and cognitive processes they are to master

REFERENCE

Bransford, J.D. and M.K. Johnson. "Contextual Prerequisites for Understanding: Some Investigations of Comprehension and Recall." *Journal of Verbal Learning and Verbal Behavior* 11 (1972): 717-726.

I gratefully acknowledge the many valuable contributions of Woody Woodford, project consultant, and the feedback of the participants and DLI staff.

COMPUTERIZED ITEM BANKING AND TEST ASSEMBLY

Lawrence S. Buck
PLANNING RESEARCH CORPORATION,
SYSTEM SERVICES

Manual test assembly and publication can be a very time consuming endeavor, especially for large scale testing programs. Hard copy item files are also very cumbersome to work with and to maintain. In order to deal with the demands of a burgeoning testing program, Planning Research Corporation (PRC) has designed, developed and implemented a comprehensive automated item banking, test generation, test publication, test scoring, and data analyses and reporting system.

The intricacies of PRC's automated Test Preparation System (TPS) will be described in detail including the test item taxonomic system, the test generation process, the publication system, and the data analyses and reporting components. PRC's automated testing system has been thoroughly field tested and placed into operational use for the November 1988 testing cycle.

BACKGROUND

Under terms of a contract with the Naval Sea Systems Command, PRC has been tasked with the lead role for the development and subsequent administration of trade-skill tests for 17 Naval shipyard trades at the journeyman level, sixteen of which have been developed to date. PRC is also responsible for effecting a major expansion of the Journeyman Naval Enlisted Classification (JNEC) testing program. PRC has been responsible for administering the JNEC testing program since 1982 when five tests were implemented. Expansion efforts began in 1986 with authorization for an additional 14 JNEC tests. By the end of 1989, PRC will have a total of 19 JNEC tests in operational use. (A total of 12 JNEC tests were administered in November 1988.)

Both the shipyard and JNEC tests consist of multiple-choice test items (four and five choice items). With an item file of approximately 300 items for each test, PRC is tasked with maintaining close to 10,000 test items. By the end of 1989, PRC will be responsible for administering some 36 tests of 100 to 150 questions each. The JNEC tests are administered biannually (May and November), while a schedule for administration of the shipyard tests has not yet been finalized. The administration requirements for the shipyard tests will, however, far surpass those for the JNEC tests.

In order to effectively and efficiently handle the rapidly increasing demands of its testing programs, PRC was faced with two options, either significantly increase the size of the testing staff (currently five employees) or automate the testing process. As might be expected, the decision was made to automate.

THE ITEM BANK

The foundation of PRC's automated TPS is the test item database. The item database, in addition to item text and alternatives, consists of an item coding system and item statistics. With respect to the item coding system, there are a number of variables that must be considered when selecting items for a test. For example, each item is assigned a unique item identifier, allocated to a specific domain (subject-matter area), and given an item usage code. Each item is also

coded to indicate whether it is a clue or inverse to another item(s), contains graphics, and whether the item's alternatives may be scrambled. (Numerical alternatives such as 10, 20, 40, 100, are placed in ascending or descending order and are not subject to scrambling.) (An item's alternatives may need to be scrambled to ensure a proper mix of correct answer locations in the printed examination.) Other encoded information addresses the specific reference from which the item was derived and the knowledge, skill or ability (KSA), that the item was written for. Reference information is important for documenting the correct answer as well as for ensuring currency of the item. The linkage to specific KSAs provides validation documentation and a means of tracking each item to the duty and task(s) that the item represents.

The most time consuming aspect of the TPS is the initial entry of test items into the database. Prior to entering the item itself, specific test configuration information identifying the test by name and number, the number of test items, and the number of domains is entered. Each test domain is then identified and the standard number of items to be allocated to each domain assigned. The data entry operator is led through this process by specific menus appearing on the CRT screen.

The next step in the process is the entry of the item's particular encoded characteristics. These characteristics include a usage indicator, the correct answer, graphics and answer sequence flags, and position requirements and item objectives. (Position requirements and item objectives are generalized and specific KSA statements respectively.) The entry of the item's text and alternatives follows, again in accordance with user-friendly menus.

Once the item codes, text, and alternatives have been entered for all the items associated with a specific test, the items' statistical histories are then entered. (If so desired, the entry of item statistical data could precede the text entry.) Summary statistics common to classical test generation such as difficulty measures and discrimination indices are entered into the TPS along with the upper/lower item response split, item number, and the date of administration. (For item selection purposes, the TPS averages each item's statistics for the three most current administrations.) The item's status with respect to whether it is appropriate for operational use or in need of further validation is also entered. The total entry process, including proofing and corrections, for a complete test item file (approximately 300 items) requires from 2 $\frac{1}{2}$ to 3 days.

The final product is an item database consisting of item codes and identifiers, item text, and item statistics. The need for voluminous hard copy item files is, for all practical purposes, eliminated, as is the need for manual posting of item analysis data. The processes of reviewing or searching for items in the TPS database is a great improvement over the reliance on item cards and manual sorts.

AUTOMATED EXAM GENERATION

The primary objective for the TPS is the automation of the examination generation process with a resulting increase in efficiency and accuracy. The first step in the process is to enter the desired test specifications addressing the domains to be included in the test, the number of items in each domain, and the desired test and item statistical characteristics. The test and item specifications are gleaned from a perusal of a TPS generated printout of items existing in the item database for the test in question. The item listing provides the basis for determining an appropriate mean difficulty level for each group of items and for the test as a

whole, the number of items to be selected from each item difficulty range (e.g., 20-29, 30-39, 50-59, etc.) for each domain, and an appropriate range for the discrimination index.

Once the desired examination specifications have been entered, the TPS proceeds to process the items within each domain into groups with difficulty and discrimination indices within the specified limits, and acceptable prior administration dates. Item selection then proceeds based on item availability as determined by each item's statistical properties, its relationship to other items, and a random factor designed to ensure that each item has an equal chance of being selected. If sufficient numbers of items are not available within a specified difficulty range, the range is expanded in either direction until sufficient items are located.

An important consideration in the preparation of tests is the capability of generating parallel tests. The TPS is capable of simultaneously developing up to three parallel exams in accordance with a specified minimum degree of uniqueness. The system is designed to determine if the desired degree of uniqueness is possible from the existing item file and then to aim for the highest degree of uniqueness possible. Partial tests consisting of sets of items from selected subject-matter areas (domains) can be generated on demand. The TPS also provides for the appending of validation items (items in need of further trial testing) to an operational test and for separate scoring and data analyses of these items.

Whereas exam items are generally selected by domain and grouped by domain, the TPS has the capability of scrambling the order of the items in an exam. The order of an item's alternatives may also be scrambled where desirable and permissible. Other editing options permit the replacement of items with items of similar statistical properties.

The automated test generation process significantly reduces the amount of time required to prepare tests in comparison with manual test generation. A trial generation of two exams of 120 items each, including the entry of test specifications, was accomplished in less than four hours. Manually, the process would have required from one to two weeks per test. Automated test generation also results in more accurate tests since human interaction is considerably lessened.

EXAM PUBLICATION

The examination publishing facet of the TPS provides for a direct link to the item database for publication of those items selected in the exam generation phase. Assuming the accuracy of the items in the database, this provides for increased exam accuracy and a significant decrease in proofreading requirements. Once the basic form of the exam has been determined, the exam is passed to the publishing system as an ASCII file. The word processor operator then incorporates graphics, formats, paginates, and prints the exam. The printed exam is then ready for reproduction and distribution.

The TPS is currently designed to interface with a separate publishing system which cannot import graphic images. Test items containing graphics are not passed to the publishing system in the form in which they are to be printed, rather they are passed as identifiers and instructions. When a graphic item is encountered during publication, the operator uses the identifier to locate the item in a special graphics file and merges the graphics with the test file passed by the TPS.

The entire exam publication process requires only a fraction of the time required for manual test generation and publishing since text entry is nearly eliminated and proofing commensurately reduced. The published test items in a TPS generated exam are absolutely faithful to the master test in the database since no transcription is required.

EXAM SCORING, ANALYSIS, AND DOCUMENTATION

The TPS scoring module employs a Sentry 3000 automatic scanner to read answer sheet responses and provide direct entry of individual exam results into the analysis system. The scanner is capable of processing 2-3 answer sheets per minute as compared to 10-20 minutes per answer sheet for hand scoring and manual data entry. The scanner is much more accurate than hand scoring and allows pertinent information such as the examinee's name, activity, and other background variables to be directly recorded as each form is processed. The use of the scanner in conjunction with the TPS allows for accurate and consistent examination grading in a fraction of the time formerly required. This enables prompt forwarding of examination results and nearly eliminates the result adjustments that often plague manual systems.

A variety of statistical analyses and reports essential for test documentation and validation are provided by the TPS. Item statistics necessary for item evaluation and for test item selection during the exam generation process are developed and recorded during the analysis phase. Analysis is initiated when an examination profile, the data file created by the answer sheet scanner for that exam, and the identity of the statistical reports desired are passed to the analysis module. The selection of statistical reports is menu-driven, allowing for selection of specific reports from among those available. Report generation and item and exam analyses are then performed automatically. Processing for a 120-item exam typically takes from 10-20 minutes depending on the reports requested.

Among the statistical analyses provided are: difficulty indices (p-values), discrimination indices (r_{bis} & r_{pbis}), reliability estimates, upper/lower group item analyses, the exam mean, and the exam standard deviation. Individual item analysis statistics are placed in a data file which is used by the TPS to update the administration statistics for each item in the file. The TPS then uses the weighted average of the statistics derived from the three most current administrations when selecting items during the exam generation process.

Numerous documentation and results reports are automatically generated by the TPS. Such reports include answer keys, domain maps, unit and individual profile analyses, and unit performance comparisons. Compilations of results both fragmented and total are also produced. In addition, the TPS system interacts with a Honeywell minicomputer to prepare individual results letters for transmittal to the separate commands.

SYSTEM HARDWARE AND SOFTWARE REQUIREMENTS

The TPS is a microcomputer based application designed to run on IBM PC/XT/AT or 100% compatible computers. MS-DOS 3.2 or later is recommended although MS-DOS 2.0 or later is acceptable. For our purposes, a minimum of 20 megabytes of disk storage is used. A floppy drive or provision for some type of removable media is required for interfacing with other systems such as the publishing system used by PRC. Currently, PRC uses the Xerox Star with PC

After thorough field testing, the TPS was used successfully to generate two of the 12 operational JNEC tests for the November testing cycle. The TPS produced tests were generated and sent to publishing in a fraction of the time normally required for manual test generation and publication. As more test item files are entered into the system, it will be increasingly relied on for generation and publication of operational tests. As operational use of the tests developed for the Naval shipyards comes to pass, the TPS will be essential in allowing PRC to meet the shipyard's needs without a dramatic increase in staffing.

The potential for further significant decreases in processing time exist as the TPS can be upgraded to accommodate scanning of items and graphics and desk top publishing with microcomputers. The TPS has proven to be sufficiently accurate, flexible, and efficient to bode well for its increased use in the future.

Test-Item Readability: A Final Report
R. Eric Duncan, Captain, USAF
Air Force Military Personnel Center

Introduction

In 1981, Duncan proposed a model of test-item readability which incorporated item, examinee, and environmental characteristics. These characteristics interact to produce the readability level of an item. Duncan (1981), however, lacked a solid theoretical framework from which to relate the semantic and syntactic components of an item to its readability. To correct that deficiency Duncan adapted Kintsch's (1974) propositional approach to text-based readability to multiple choice test items. This paper will briefly describe this adapted approach, describe the most critical syntactic and semantic variables which predict item readability, and present the final results in the attempt to predict item readability.

Semantic Variables and Kintsch's Propositional Theory

The variables described here can be grouped into four distinct areas: semantic variables, syntactic variables, a cognitive load variable, and measures of prior knowledge. The semantic variables include propositional density, operator density, argument density, and propositional level. The variables in this area are directly related to Kintsch's (1974) propositional approach to the description of memory in semantic memory and so a brief explanation of that theory is in order.

Kintsch's (1974) semantic approach to reading and processing textual material focuses on the proposition. Kintsch and Keenan (1973) point out that sentences read from text are not stored verbatim, but rather as propositions. Propositions are word concepts combined to form a logical set of lexical items and contain a relation (usually a verb) and » arguments (nouns, adjectives, pronouns). These propositions are put together in a logical manner, establishing a representation of the text, known as a text base, in memory. A text base is simply "an ordered list of propositions" (Kintsch, 1974, p. 13). To obtain this list of propositions, the text must be semantically analyzed. Propositions, which include relation and argument(s), are then abstracted beginning with the first sentence.

Before describing the propositional construction process, relations and arguments need to be described. A relation is a word concept (not necessarily a word) which describes some action or state of being and normally appears as a verb, adjective, adverb, or noun.

John sleeps (SLEEP, JOHN) (1)

Mary bakes a cake (BAKE, MARY, CAKE) (2)

In example (1) from Kintsch (1974), SLEEP, a verb, is the relation describing some action that is being performed by the argument, JOHN. Fillmore (1971) established semantic rules for arguments. Arguments must be an agent, experiencer, instrument, object, source or goal and are structured in that order of importance. In example (2), MARY is the agent that BAKE(s) the object CAKE. Propositions can serve as arguments for other propositions, as well. In example (3),

If Mary trusts John, (TRUST, MARY, JOHN) = a (3)

She is a fool (FOOL, MARY) = b

(CONDITION: If, a,b) = c

the proposition "c" has propositions embedded in it as arguments. This function is important when building text bases, since it is more economical and requires less memory space than recreating new propositions that had been processed earlier.

Levels and Operators

Two other important features in the derivation of a text base are presented in example (3). The first feature is that of the level of propositions. "A proposition is said to be subordinate to another if it contains an argument that also appears in the first proposition" (Kintsch, et al., 1975). Subordination can occur, as is most often the case, immediately after the superordinate proposition in the text base, or can occur much later in the text base, as in the case of propositions being used as arguments for other propositions. In example (3), subordination is indicated by the indentation of propositions "b" and "c". Indentation of propositions is a convention established by Kintsch and was used in creating item protocols.

The second important feature shown in example (3) is that of operators. Operators are mechanisms which require inference on the part of the reader. Simply, operators do not state an explicit relationship among propositions, but rather challenge the reader to obtain an implicit meaning from preceding information. Operators can also be described as taxing memory space and requiring memory searches to determine what and how to manipulate previously encountered propositions. They include: Causality, Contradiction, Part, Time, Location, Condition, Conjunction, and Purpose. The operator would normally appear in the listing of the text base as presented in the "c" proposition. One additional operator (MATCH) has been created that was not included in Kintsch's operators. This operator identifies the cognitive operation of matching the propositions in the item alternatives to those propositions stored in memory.

Empirical Evidence in Support of Kintsch's Theory

Now that the basic components of Kintsch's theory have been described, this section provides the empirical evidence which supports the contention that test items are broken down into logical semantic units (propositions) before storage in memory.

Kintsch and Monk (1972) demonstrated that experimental subjects stored text material in the same manner, regardless of the syntactic complexity of the text. They found that the more syntactically complex paragraphs took longer to read but that there was no significant differences in the number of propositions recalled. Kintsch and Monk suggested that text is not represented syntactically in memory, but rather that it is represented semantically in propositional form. This evidence supports the contention that text is parsed into propositions and is stored in semantic memory.

Kintsch and Keenan (1973) examined the effect on reading time and recall of the number of propositions and the level in text of the propositions. The length of sentences (total number of words) and the number of propositions they contained were covaried. This approach is better known as propositional density, i.e., number of propositions/number of words. The levels of propositions were also varied. Kintsch and Keenan found that, if reading time was unlimited, propositional density significantly affected recall rates. They also demonstrated that superordinate propositions were recalled better than subordinate propositions. This finding was later supported by Kintsch, et al. (1975). This result can be more easily explained by referring back to example (3). Proposition "a" is a superordinate proposition while propositions "b" and "c" are subordinate to "a". Kintsch and Keenan, and Kintsch, et al. have shown that proposition "a" has greater probability of recall than propositions "b" or "c".

In addition to supporting the results of Kintsch and Monk (1972) and Kintsch and Keenan (1973), Kintsch, et al., (1975) also examined

the effects that the number of different word concepts (arguments) would have on recall. Results indicated that the more frequently a word concept (argument) is repeated in the text, the better it is remembered. The authors also showed that, as the number of different arguments increases, reading time increases and recall decreases. Kintsch, et al. suggested that the history paragraphs may be easier than the science paragraphs they used because they contain propositions that "are already part of the subjects' general knowledge" (p. 209).

Variables of Interest

Experimental evidence (Duncan, 1985) has shown that propositional density (# of propositions/# of words), operator density (# of operators/# of propositions), argument density (# different arguments/# of propositions), and propositional level contribute to reading comprehension. The hypothesized relationship between propositional density (PD) and reading comprehension scores in test items was that as PD increased (more propositions per word), the reading score necessary to understand the item would also increase. This is also true for operator density (OD), argument density (AD), and propositional level (PL).

Syntactic variables from Duncan (1981) must also be included in the estimation of test-item readability. Centerembeddedness, a modifying word or phrase located between the subject and predicate of an item, has been shown to make text less comprehensible when present (Lambert and Siegel, 1974). A modifying phrase that precedes the subject of an item is known as a left-branched phrase. A phrase which follows the predicate of an item is known as a right-branched phrase. Schwartz, et al., (1970) showed that left-branched phrases reduced the comprehensibility of text while right-branched phrases had no appreciable effect on comprehensibility. These three variables, then, constitute the syntactic element of test-item readability.

There are two other elements that contribute to the prediction of item readability: (1) measures of prior knowledge, and (2) cognitive load. Kintsch and Vipond (1977) suggest that prior knowledge could enhance a person's ability to extract meaning from text. They indicate that the best method to assess prior knowledge is with a vocabulary test containing words used in the text, material being read. This direct method, in the present study, is impossible to apply since all test and subject data had been taken from historical data files. Substitute measures include jargon and uncommon words. The number of uncommon words (UW) is determined by comparing all words in an item to the *Common Word List* compiled by Kincaid, et al., (1980). This list contains the 20,000 most common words used by enlisted Navy personnel with a 9th grade reading ability. After comparison, those words not appearing in the list are compared to the text reference material to determine if they are sufficiently explained. If there is no explanation or definition given, the word(s) are counted as uncommon. The second variable used to assess prior knowledge is jargon. Jargon is based on the judgement of the item raters'. Jargon words, such as CBPO, grade, and MAJCOM are not common to the general public and are specific to Air Force personnel. The jargon variable is expected to enhance readability since the use of jargon is commonly used in the military to communicate frequently complex names or phrases in a succinct manner. The variable used to measure cognitive load is Bloom's Taxonomical Level (Bloom, et al., 1956). This variable indicates the cognitive activity necessary to read and answer a test question and includes the following levels: (1) rote memory, (2) comprehension, (3) application, (4) analysis, (5) synthesis, and (6)

evaluation. Raters will evaluate each item on this variable and indicate the appropriate level.

The variables of interest, then include four semantic variables adapted from Kintsch's propositional approach, three syntactic variables, two measures of prior knowledge, and a variable to assess cognitive load.

Experimental Approach

Two hundred multiple-choice items from a commonly used Air Force test were evaluated on the variables of interest by seven trained raters in a counter-balanced fashion. The data was reviewed by the author to insure accuracy. Two thousand examinees were randomly selected and their Armed Services Vocational Aptitude Battery scores were used to determine each examinee's reading grade level (RGL). This was accomplished using regression equations provided by Madden and Tupes (1966). The RGL of the 25th percentile of all examinees answering each item correctly was assigned as the RGL for that item.

The RGL and p -value for each item were used as the dependent variables in further analyses. Regression analyses were performed to determine if statistically significant equations to predict test-item readability and item difficulty could be produced. Two regression approaches were attempted - all possible subsets regression and stepwise multiple linear regression. The independent variables were the item variables previously described.

Results and Discussion

When examining the descriptive statistics for both low and high-ability test items, some very distinct differences appear. The 25th percentile reading grade level (RGL) for the items in the low and high-ability tests was significantly different ($t=63.79$, $p<.001$) with a mean for the low-ability items of 8.916 and a mean of 10.016 for the high-ability items. This difference was anticipated, in that, by definition low-ability items should have lower RGL's than higher-ability items. While there were no significant differences between high and low-ability items on non-semantic variables, interesting trends were present. For the high-ability test, right-branched phrases occurred somewhat more frequently (81 vs. 76 percent); there were fewer left-branched phrases (21 vs. 22 percent), and fewer uncommon words (14 vs. 20 percent) were present than in the low-ability test. As expected, low-ability test items were more frequently written at a lower level in Bloom's taxonomy than high-ability test items. Propositional, operator, and argument densities and propositional level were somewhat more densely packed on low-ability items than on high-ability items, but not significantly so. This, however, can be seen in the fact that low-ability items had to present more information in a syntactically simpler fashion. The reading grade level (RGL) of the text from which low-ability items were taken was also somewhat higher than for high-ability items. This resulted in lower-ability examinees taking items gleaned from higher RGL text. As noted earlier, though, test construction personnel were able to minimize the low-ability items' RGL. A factor analysis was conducted on both tests using the principle axis with iteration method. The factor pattern matrix was rotated orthogonally. Using a scree test, only the first two factors of the low-ability test were determined to be significant with eigenvalues of 4.394 (20.7 percent of the variance) and 1.301 (6.1 percent of the variance) respectively. The same method, used with the high-ability test, revealed two

significant factors with eigen values of 5.339 (24.1 percent of the variance) and 1.661 (7.5 percent of the variance) respectively. The results supported the unidimensionality assumption. Intraclass correlations revealed that raters were able to consistently rate the semantic variables showing that training given to the raters was effective.

Regression analyses were performed using both item difficulty (p -value) and item RGL for low-ability items, high-ability items, and all items combined. Looking at item RGL, the multiple regression obtained from a calibration group of 102 low and high-ability items resulted in a multiple R of .462 and an R^2 of .213 ($F_{9,77} = 5.776$, $p < .01$). However, a plot of the standardized residuals revealed the positive residuals came from the high-ability test and the negative residuals came from the low-ability test. Such a finding suggested the need to examine the low- and high-ability items separately. The regression equation for item RGL on the low-ability test shown below was insignificant ($F_{9,0} = 1.155$, $p < .05$).

Item RGL = $8.9246 - (.01192 * \text{right-branching}) + (.02091 * \text{Bloom's Taxonomical Level}) - (.01162 * \text{left-branching}) - (.00296 * \text{text RGL}) + (.05037 * \text{argument density})$

For item p -value as the dependent variable and using low-ability items, the equation produced a significant multiple R of .297 and an R^2 of .088 ($F_{9,1} = 2.96$, $p < .05$).

Item p -value = $90.2986 - (1.6738 * \text{text RGL}) - (19.0181 * \text{propositional density}) - (1.3964 * \text{jargon})$

This equation suggests that, for low-ability items, items become easier to answer as the RGL of the text from which the item was taken decreases. This corresponds to the decrease in propositional density as items became easier.

For high-ability items, the regression equation for item RGL was not significant. However, for item p -value the equation produced a multiple R of .287 and an R^2 of .081 ($F_{9,3} = 2.70$, $p < .05$).

Item p -value = $99.5153 - (41.6326 * \text{propositional density}) - (8.0535 * \text{center embeddedness}) - (33.6494 * \text{operator density})$

Like the low-ability items, the easier an item is, the fewer propositions/words and operators/propositions exist.

Conclusion

For item RGL only, the combined regression equation was significant. However, this equation, while not presented here, loaded a predominance of semantic variables. This suggests that when developing a heterogeneous group of items, propositional level, operator and argument densities, and text RGL are the best to establish an item's RGL. It also contradicts conventional wisdom that using text RGL equations is suitable for multiple-choice test items. Upon closer examination of the regression equations and other analyses, it seems that low-ability examinees are confused by too much semantic information and have difficulty processing a high RGL item. For these individuals, simple syntax and even more simple semantics (few inferences) aid in their understanding of an item. However, the opposite is true for high-ability examinees. Syntax has very little to do with understanding an item. These examinees seemingly read and understand study material more semantically. The use of high levels of semantics in items (more operators, deeper propositional level, etc.) seems warranted with these examinees.

Yet, when item difficulty is considered, reduced levels of propositional variables makes the items easier to answer. It then

appears to be a tradeoff -- difficulty or readability. It is not as simple as a tradeoff as there was no significant relationship between difficulty and RGL found in this study. When developing multiple-choice test items, both difficulty and readability must be considered equally.

Bibliography

- Bloom, B.S., M.D. Engelhart, G.J. Furst, W.H. Hill, and D.R. Krathwohl, 1956. Taxonomy of educational objectives (subtitle: The classification of educational goals): *Handbook 1, The cognitive domain*. New York: David McKay Company, Inc. Portions reprinted in G.H. Bracht, K.D. Hopkins, and J.C. Stanley (eds.). 1972. *Perspectives in educational and psychological measurement*. Englewood Cliffs, N.J.: Prentice-Hall, Inc., Selection 12.
- Duncan, R. Eric. *A Theory and Model of Item Readability*. Published in the Minutes of the 23rd Annual Conference of the Military Testing Association, 1981.
- Duncan, R. Eric. *A Theory and Model of Item Readability*. Published in the Minutes of the 25th Annual Conference of the Military Testing Association, 1985.
- Duncan, Robert E. "Theoretically Based Test Item Readability: An Approach to Estimating the Degree to which an Item Can Be Understood and Answered Correctly." (1986). *Dissertation Abstracts International*.
- Filmore, C. J. Some problems in case grammar. Cited in Kintsch, W., *The Representation of Meaning in Memory*. Hillsdale, N.J.: L. E. Earlbaum Associates, 1974.
- Kincaid, J.P., Aagard, James A., and O'Hara, John W. *Development and Test of a Computer Readability Editing System (CRES)*. Training Analysis and Evaluations Group Report No. 83, March, 1980. TAEG; Orlando, Florida, 321813.
- Kintsch, W. *The representation of meaning in memory*. Hillsdale, N.J.: L. E. Earlbaum Associates, 1974.
- Kintsch, W. and Keenan, J.M. Reading rate and retention as a function of the number of propositions in the base structure of sentences. *Cognitive Psychology*, 1973, 5, 257-274.
- Kintsch, W., Kozminsky, E., Strobry, W. J., McKoon, G., and Keenan, J. M. Comprehension and recall of text as a function of content variables. *Journal of Verbal Learning and Verbal Behavior*, 1975, 14, 196-214.
- Kintsch, W. and Monk, D. Storage of complex information in memory: Some implications of the speed with which inferences can be made. *Journal of Experimental Psychology*, 1972, 94(1), 25-32.
- Kintsch, W. and Vipond, D. Reading Comprehension and Readability in Educational Practice and Psychological Theory. In Lambert, J.V. and Siegel, A.I. Psycholinguistic determinants of readability. In A.I. Siegel and J.R. Burkett (Eds.) *Application of structure-of-intellect and psycholinguistic concepts to reading comprehensibility measurement*. AFHRL-TR-74-49, Lowry AFB, Colo., 1974.
- Madden, H.L. and Tupes, E.C. *Estimating reading ability level from the AGE General Aptitude Index*. Lackland Air Force Base, Texas: Personnel Research Laboratory, Aerospace Medical Division, PRL-TR-66-1, AD-632 182, February 1966.
- Schwartz, D., Sparkman, J.P., and Deese, J. The process of understanding and judgments of comprehensibility. *Journal of Verbal Learning and Verbal Behavior*, 1970, 9, 87-93.

Prediction of Sensitive Compartmented Information (SCI)
Access Using Personality Tests

LeRoy A. Stone, Ph.D., ABFF, ABPP
Department of Defense, Ft. George G. Meade, MD

Security clearances, as granted by the US Government, are of several different types and levels. Although most persons are rather familiar with the concepts, secret and top-secret clearances, very few have ever encountered the Sensitive Compartmented Information (SCI) clearance concept. SCI access is, for all practical purposes, the highest major security clearance level granted to any fairly large group of people by the government. The decision to grant access to the SCI clearance level is determined by guidelines specified in a Directive issued by the Director of Central Intelligence which is best known as the DCID 1/14 (see Director of Central Intelligence, 1984). Although such information may be very difficult to estimate and obtain, Washington Post reporters (see the June 8, 1986 issue, A Section, page 18) have stated that as of March, 1985 there were 98,715 persons (civilian and military) within the Department of Defense (DoD) and another 9,576 in industry who held SCI access. In order to obtain SCI access, individuals may/must undergo what might be considered as extremely rigorous screening or vetting which employs psychological testing and interviews, polygraph examinations, extensive and depth background investigations and records checks. Reinvestigations are supposedly routinely conducted every five years for those who hold this particular clearance.

Although the psychological testing instruments which may be involved in evaluating persons as part of the clearance screening/vetting process may differ from one agency or organization to another, there are at least two popular personality measuring tests which have been regarded as appropriate for this particular kind of purpose or situation. Historically, the Minnesota Multiphasic Personality Inventory (MMPI) immediately comes to mind as perhaps useful for this type of assessment situation. In more recent years, the Millon Clinical Multiaxial Inventory (MCMI) is another testing instrument which has had some use in similar type assessment situations. The question of whether these two very popular, frequently utilized, personality-measuring instruments do provide psychometric information which can be used in prediction of whether eventual high-level security clearances (SCI access, in particular) will or will not be subsequently granted was the raison d'être for this investigation. The research reported in this paper is of the kind that was recently suggested, by a Congressional Committee (House of Representatives, 1987), be conducted. They urged that "research attention be given to . . . personality-based profiling . . . as an element to be considered in the screening process (pages 7-81)."

Study 1

The MMPI was administered to a sample of 102 contractor employees (these were individuals who were employees of companies/corporations which were providing services to the DoD) who were being psychologically evaluated and processed for possible granting of SCI access; 61 were eventually and finally successful in obtaining SCI access, 41 were not. In this particular sample of 102, their ages ranged from 22 to 60 years ($M = 36$, $SD = 10$), 17% were female and 83% were male, average education in years was 15.5 years ($SD = 2.7$). Approximately 25-30 different contractor firms were represented as sponsors for this sample. Most of the employees were from either the West or East Coast regions. Based upon their MMPI clinical scales scores, this sample appeared to be rather representative of the general population as their MMPI scores seem to be rather close to recognized recent norm values (e.g., see Colligan, Osborne, Swenson & Offord, 1983). This particular sample can be regarded as being somewhat unusual in that the rate for nonobtaining access/clearance can be considered to be unusually high. However, this sample can be regarded as well suited for a validity testing of the MMPI with respect to the question of eventually obtaining/nonobtaining SCI access.

In actuality, the criterion or validity variables were three in number. The first of the three was whether the interviewing clinical psychologists gave positive or negative recommendations for the individuals involved. The second criterion variable was whether the chief of the clinical psychologists also gave positive or negative recommendations following his review of all pertinent information. Only in a couple of cases did the chief psychologist offer a different recommendation than did the interviewing psychologists; in all such cases he made a negative recommendation whereas the interviewing psychologists had given positive recommendations. Also, it should be noted that the only real differences between the second criterion variable and the third (which in fact was the final adjudication decision, made by the Office of Security, as to whether the individual would or would not be granted SCI access) were that the decisions made in the third were generally more stringent. Correlations between all three criterion variables were high. The correlation coefficient for between the interviewing psychologists' recommendations and the chief psychologist's review recommendations was .96; between the interviewing psychologists' recommendations and whether SCI access was granted or not was .836; and between the chief psychologist's review recommendations and whether SCI access was granted was .837. It should be noted that none of the interviewing clinical psychologists (nor was anyone else in the investigation/adjudication chain) were aware that this investigation was being conducted while the data were being collected.

To analyze the mean differences between the contractor employees judged Unfavorable versus Favorable by the

interviewing psychologists, two-tailed t-tests were calculated, based on T-scores from the basic MMPI scales. The results of these comparisons reveal that only for the K, Hs, and the Sc scales were the differences between means perhaps even close to being statistically significant; these three differences between means could only be regarded as being significant at the .10 level and not significant at the .05 level. Also, most of the differences were in what might be considered as an unusual direction; higher psychopathology was indicated with the favorably judged group than was observed for the unfavorably judged group with two of the three scales. Such a directional difference is exactly opposite to what one would predict. It seems reasonably safe to conclude that all of the tested differences could have easily occurred, simply on the basis of chance alone.

To analyze the means differences between the contractor employees judged Unfavorable versus Favorable by the chief psychologist upon case review, two-tailed t-tests were calculated based on the T-scores from the basic MMPI scales. The results of these comparisons reveal that only for the K, Hs, and the Si scales were the differences between means perhaps even close to being statistically significant. The difference between Si means could only be regarded as significant at the .10 probability level, whereas with the K and the Hs scales the differences could be regarded as significant at the .05 level but not at the .02 level. Again, only with the Si scale difference did the direction of the difference make any sense, very slightly more psychopathology indicated for the Unfavorable group than with the Favorable group. Again, it seems reasonably safe to conclude that all of the tested differences could have easily occurred, simply on the basis of chance.

To analyze the differences between means of the contractor employees judged Unfavorable versus Favorable by the complete SCl access adjudication process, two-tailed t-tests were calculated, based on the scores from the basic MMPI scales. Inspection of the computed t-tests revealed that only for the Sc scale was the difference between means perhaps even close to being statistically significant; the difference could only be regarded as being significant at the .10 probability level. Since this one noted difference was just one of the 13 t-tests computed with these data, then the most likely explanation for the size of the difference was that it occurred simply on the basis of chance alone.

It should be noted that the t-test statistic was not the only one used to explore whether MMPI scores were systematically related to the three criterion variables. Multiple (two-group) discriminant functions were also computed with each of the criterion variables and included all 13 studied MMPI variables. All three discriminant functions were found to be statistically nonsignificant (p-levels were not even remotely close to any frequently utilized alpha levels).

It seems rather safe to conclude that any differences noted between means (for the Favorable vs. Unfavorable groupings) were simply based only on chance alone. In other words, there would appear to be no association noted on the MMPI with respect to whether contractor employees, being psychologically evaluated and finally evaluated by SCI access adjudication process, were later regarded favorably or unfavorably in the psychological evaluations or more finally by the Office of Security. Psychopathology, as measured by the MMPI, seems to be independent of psychological evaluations (at least with contractor employees) and of SCI adjudication decisions.

Study II

The MCMI was administered to a sample of 117 male contractor employees who were being psychologically processed for being possibly granted SCI access. With the exception that this sample was composed of all males, it resembled very closely the sample which was used in Study I of this report. The rate for successfully obtaining SCI access was however a bit higher for this group than noted in Study I. With this sample of 117 employees 88 (or 75%) were successful in finally obtaining SCI access status.

Exactly the same statistical analysis procedures which were used in Study I were also utilized here in Study II. It was decided that separate analyses for the evaluating psychologists' recommendations were not to be carried out, but rather only the final SCI access adjudication decision would be regarded as the single, only criterion variable. To analyze the differences between MCMI scales means of the contractor employees judged Unfavorable ($n = 29$) versus Favorable ($n = 88$) by the complete SCI access adjudication process, two-tailed t-tests were calculated, based on the scores from the 20 MCMI scales. Inspection of the computed t-tests revealed not one of them significant at any generally used alpha level setting. The largest t value noted was 1.76 and, with 115 degrees of freedom, this was not significant at the .05 level. A two-group discriminant function was also computed and a 'less than unity' F-test value (dfs 20 and 96) was obtained, indicating that there was no systematic or reliable relationship between the MCMI data and the two-group adjudication decisions regarding the granting of SCI access.

Discussion of Results of Studies I and II

It is interesting to note that neither the MMPI nor the MCMI result in psychometric information which is predictively associated with success or failure in obtaining SCI access status, at least not with the defense contractor employees in the two samples studied. Although there has not been a great deal of reported research dealing with this particular kind of problem, these current findings are actually consistent with

results of a couple earlier reported investigations. About a decade and a half ago, within the same intelligence agency a rather similar investigation was conducted; a study of the relationship of MMPI scales scores (and some other demographic and psychometric information) of applicants for employment was completed and they essentially found the same kind of results. Namely, with male applicants (and it should be noted that in the currently reported 'Study I', 83% of the studied contractor employee sample were males) they found that only two MMPI scales (the F and K scales) might be associated with interviewing psychologists' favorable-unfavorable recommendations. They concluded that there were no substantive mean differences between Fit and UnFit male Ss on the clinical MMPI scales. More recently, a study (Hibler, Kolski, & Chapman, 1988) was reported which also had some findings which were perhaps similar in fact to the present findings but which appear to have been confusingly interpreted by the principal investigator of that study. They reported (based on a study of 122 active-duty military members who were involved with SCI access) that "a comparison of individuals with elevated MMPI clinical scales (>70T) to those with non-elevated MMPI clinical scales, suggested that elevated MMPIs do not differentiate those receiving a recommendation to deny access from those retaining access [page 2]." This kind of finding seemed to be about equally true with their MCMI data scores information. Rather surprisingly, they then concluded, based on their research data pertaining to the matter of maintaining or obtaining SCI access, that "the objective personality tests, in particular, were most useful for identifying significant pathology [page 5]." In actuality, their results rather cleanly suggested just the opposite conclusion.

With all these findings (including those presently reported), it would seem that it would be most unwise to attempt to build any type of paradigm or scheme for predicting whether applicants or candidates would be later favorably or unfavorably adjudicated for SCI access and base such a paradigm on MMPI or MCMI scales score information. What appears to be the major problem with using this kind of particular psychometric information is perhaps due to response set or style problems and difficulties. The investigator is very familiar with this kind of psychometric information used as part of psychological evaluations conducted in the SCI access adjudication process. It has been repeatedly observed that some rather potent personality-test kinds of response styles very frequently and significantly influence responses of applicants/candidates who are being considered for possible SCI access. It has been observed that a good percentage of such individuals seemingly attempt to minimize psychopathology reporting and this kind of minimization seems to be relatively independent of true psychopathology. Also, it is not at all unusual to encounter other individuals who attempt to be more-than-candid or to be more than usually honest in their self-reporting of psychopathology and its possibly related behavioral correlates.

For example, it is not unusual to find that some individuals, who are sometimes noted to be associated with particular religious faiths, tend to exhibit a more-than-usually-encountered level of honesty in their own self-descriptions (particularly on psychometric inventories/questionnaires). This kind of unexpected degree of honesty, with respect to self-description, seems to also be somewhat unrelated to actual or true psychopathology. Because of these kinds of sources of measurement error, it becomes very difficult to conclude that low scores on a MMPI or MCMI type instrument actually reflect low levels of psychopathology or that high scores on these kinds of instruments actually reflect high levels of psychopathology.

It should be noted that the research results in this reported investigation should not be interpreted to mean that the MMPI or the MCMI should not be clinically used as psychological tests in the psychological evaluation process involved in SCI adjudication matters. The results though do indicate that MMPI and MCMI scores, based upon the basic or major scales of these two instruments, should not be used in any fashion which would suggest that they were the sole basis for any favorable or adverse determination and/or for prediction purposes in the SCI adjudication process.

References

Colligan, R. C., Osborne, D., Swenson, W. M., & Offord, K. F. (1983). The MMPI: A contemporary normative study. New York: Praeger.

Director of Central Intelligence (1984). Director of Central Intelligence Directive No. 1/14. Washington, DC: Author.

Hibler, R. J., Kolski, A. R., & Chapman, R. K. (1988). Pathological indices in personnel with high security clearances. Paper read at the 12th Symposium, Psychology in the Department of Defense. Colorado Springs, Colorado.

House of Representatives (1987). Report by the Permanent Select Committee on Intelligence. (Report 100-5). Washington, DC: U.S. Government Printing Office.

Disclaimer

The opinions and findings of this paper are those of the author and do not necessarily reflect those of the Department of Defense or the Federal Government.

A MICROCOMPUTER TEST BATTERY: NORMATIVE DATA AND SENSITIVITY TO MILITARY STRESSORS

Robert S. Kennedy, Dennis R. Baltzley, & Mary K. Osteen
Essex Corporation
Orlando, FL

ABSTRACT

Mental tests can provide early indication of impairment in operational performance which may be due to environmental hazards or toxic chemicals. Microcomputers can improve on paper-and pencil media because of their speed, automaticity, diversity of factors tested and other features. As with traditional approaches, suitability requirements for such test materials include satisfying metric criteria and practical factors. This paper reviews several interlocking normative studies which have yielded a menu of tests which demonstrate specific metric features: stability, task definition (stabilized reliability), reliability efficiency (reliability standardized to a three minute base), and factor diversity. The recommended short (< 15 min.) and medium (< 20 min.) batteries are available with factor loadings. Some predictive validities and other normative data are available for young adults. Validity data in the form of correlations with intelligence tests and several sensitivity studies (alcohol, drugs, sleep loss, mixed gas, simulated altitude, and chemoradiotherapy) with repeated administrations of the APTS are available. The APTS requires no special interfaces and is implemented on two off-the-shelf, fully portable and compact microcomputers, (NEC PC8201A and the Zenith ZFL-18X series) and most IBM compatible systems and is specifically designed to be used in adverse conditions.

APPLICATIONS OF THE AUTOMATED PERFORMANCE TEST SYSTEM (APTS)

The presence of environmental stressors and toxic elements in military and space environments, as well as the ordinary workplace, makes desirable the development of an assessment tool to detect subtle differences in mental acuity before such changes threaten operational efficiency. With the availability of such a tool, the performance effects of environmental stressors of interest to the military and others can be studied such as thermal extremes, hyper- or hypobaria, motion, vibration, noise, sensory deprivation or overload. From such study, exposure limits can be specified. Other applications range from screening key persons in responsible jobs (e.g., nuclear power plants) for fitness for duty and providing feedback to susceptible personnel to exploring the possibility of coping methods, adaptation, and resistance training.

Development of the APTS has followed the assumptions of classical test theory (Allen & Yen, 1979) and the empirical findings of the Performance Evaluation Test for Environmental Research (PETER) program (cf. Bittner et al., 1986). Begun by the Navy (Kennedy & Bittner, 1977), the chief outcome of the PETER program were statistical methodologies with which to evaluate tests for repeated measures applications and thirty "good" tests; mostly in paper-and-pencil modes. Over the last decade, and through the support of organizations such as NASA, National Science Foundation, and corporations in the private sector, a program of study for microcomputer implementation, mechanization, and psychometric development has been carried out. Somewhat self-consciously, the model for this program has been the APA standard for construction of educational and psychological tests (AERA, APA, & NCME, 1985) and the tenets of mental test theory (Gulliksen, 1950). Some statistical by-products specifically related to repeated measures application have been ancillary to the main purpose of this program. These include: correlated averages versus averaging correlations (Dunlap, Bittner and Jones, 1983), optimization of test length according to reliability (Dunlap, Jones, Kemery, & Kennedy, 1986), and surrogate tests to improve operational performance measurement (Kennedy, Lane, & Kuntz, 1987). The APTS psychometric development has included culling of each tests' properties to produce a menu of acceptable tests of cognition, information processing, psychomotor skill, memory, mood, and others. From this menu of tests a battery may be tailored to suit specific applications.

PSYCHOMETRIC CRITERIA

Studies of environmental and toxic agents ordinarily follow a repeated measures experimental testing protocol where each subject is his own control. Generally, tests are administered in a pre-per- and post- paradigm (PPPP or P⁴). Often the treatments which are administered are randomly administered and fully crossed over on subsequent trials. Repeated-measures designs are usually more economical and powerful than alternate approaches (Winer, 1971) and are ideally suited to experiments with small numbers of subjects. However, we believe that when test batteries are compiled insufficient attention has been paid to the special statistical requirements for this design (Bittner et al., 1986). Specifically, we suggest that not only level or asymptotic mean scores are required but more importantly, invariant retest correlations and constant variances. The latter comprise the compound symmetry requirement of the variance covariance matrix (Winer, 1971, pp. 276-277), and must be demonstrated. So far as we know, no other battery has applied such stringent criteria for qualifying tests for inclusion.

The process of qualifying a test for inclusion in the APTS is to exceed each of these metric requirements as they are successively applied. These have been described at great length in earlier reports (Bittner et al., 1986; Kennedy, Baltzley, Osteen et al., 1988) and are listed briefly in Table 1. In the APTS development work, the chief criteria that were focused upon during early battery development were "stability" and "reliability." Descriptively these criteria imply that the measured performances of individuals over sessions should be parallel and as sharply defined as possible. We recommend $r = .707$ as a lower bound for retest reliability. Later in the program of development and as more tests were shown suitable, we proposed that "better" tests were those which were uncorrelated with others and conservalational of the demands on the subject's time.

The early work, which has been reported previously in a series of publications, is available along with a demonstration disk of 30 tests (some having several versions) on an IBM compatible five-inch floppy disk from Robert S. Kennedy, Essex Corporation, 1040 Woodcock Road, Suite 227, Orlando, FL 32803. The tests are also programmed for presentation on NEC PC8201A and Zenith 181-3 PC, battery operated computers. The purpose of this paper is to synthesize and update of the progress of the dozen interlocking normative studies, to report factor and correlation estimates, as well as describe a set of studies where sensitivity of the tests to certain treatments are examined.

INTERLOCKING STUDIES AND VALIDITY CONCERNS

More than a dozen normative studies have been completed with APTS to date using the repeated-measures paradigm. A small sample of subjects (usually 25) takes the APTS for 7 to 15 sessions. The current paper presents the combined results of eight of these studies. The remaining four studies consist of two which are currently under analysis (with $N = 100$), one with insufficient replications (two trials) to be included, and one which begins a new normative effort, a vision battery of six temporal acuity tests.

Generally, the cross-correlations with the tests from these studies with the APTS show good discriminant and convergent validity. Discriminant validity is evidenced through intercorrelations which are low to moderate between tests tapping different psychomotor or cognitive abilities. For example, while the correlations between certain motor tests (e.g., Tapping) and perceptual tests (e.g., Pattern Comparison) are very low, correlations between similar cognitive tests (Grammatical Reasoning and Manikin) are $> .50$. Convergent validity is shown through high correlations between tests measuring similar constructs. For example, the three different Tapping tests correlate highly with each other and with simple Reaction Time (1-choice), and the correlations drop as the complexity of the Reaction Time tests increase (i.e., as the test becomes more cognitive). In the test development phase we compared the APTS with the original paper-and-pencil versions and found them to be interchangeable. Next, reliabilities (Task Definition) for each test were found to be high (Table 2). These findings, combined with the comparability of

results across eight studies, allows us to state with some assurance that establishment of construct validity is well under way for the APTS.

Factor Content. Although factor labelling involves an element of risk with respect to the "true" content of the factor, a synthesis of factor and correlational analyses across a series of studies suggests the following interpretation (cf. Lane & Kennedy, 1988). There are at least three important factors in the APTS tests that consistently recur in various studies (even in early trials), and a fourth factor that emerges at or around the trial at which most tests are stable. 1) Motor Speed - speed of response execution, particularly those for which the "rules" are simple and output is in part dependent on how rapidly responses can be entered. 2) Symbol Manipulation/Reasoning - involves a "generalized" ability to reason abstractly through the application of rules rather than the learning or remembering of the early rules themselves. 3) Cognitive Processing Speed - reflects the extent to which defined rules governing generation of response alternatives for a particular test have been learned through practice and can be used progressively more rapidly. 4) Response Selection Speed - the speed with which responses can be selected from the generated set of response alternatives.

The right columns of Table 2 show the relative importance of these factors for each of the tests after most tests have reached stability (generally trial 4). Since the estimates of loadings and patterns were obtained from a number of different factor analyses over a series of studies involving differing variable sets and sample sizes, and since a number of these analyses were necessarily based on relatively small numbers of subjects, the loadings are represented in terms of the patterns seen in analyses rather than in terms of absolute loadings. More complete information may be found in Kennedy, Jones et al. (1988) and Lane and Kennedy (1988).

Predictive Studies. Four holistic measures of intelligence (Wechsler Adult Intelligence Scale-Revised; American College Test; a synthetic Armed Services Vocational Aptitude Battery, and the Wonderlic) were administered to 37 subjects and compared to performance on twelve "good" APTS tests (Kennedy, Baltzley, Dunlap et al., 1988). Multiple correlations implied strong correspondence between APTS and global measures of intelligence. R^2 adjusted for shrinkage ranged from 21 to 65% common variance cross-validating previous findings (Kennedy et al., 1985). Recently the battery has been task analyzed and the tasks have been compared with the job elements necessary for space station personnel (Jeanneret, 1988).

Sensitivity Studies. The APTS has undergone empirical validation in several sensitivity studies, six of which are epitomized in Table 3. Four of the studies use each subject as (in a "placebo" condition) as his or her own control, and the two others used either cohorts (chemo) or control groups (drugs). The APTS has also been used in studies involving sleep loss, cave dwelling, exercise interventions for fatigued airline pilots.

The sensitivity studies of Table 3 are a preliminary attempt at creating a series of dose equivalent matrices. Although available on limited number of individuals and conditions, they are illustrative of our long range plan. In Table 3 we compare compare 1) three blood alcohol levels, .05, .10, .15, 2) simulated altitude at 15-20K at 23-25K; 3) motion sickness drugs, scopolamine and a combination of scopolamine and dexedrine; 4) effects of chemoradiotherapy, reported as an average decrement in treated versus untreated cohorts (donors); 5) two experimental drugs; and 6) halon gas decrement (averaged across 24 hour exposures). It may be seen that the blood alcohol levels based against placebo show an orderly loss in performance from BAC .05 through BAC .15. We suggest that this relation be used as a preliminary marker to index other comparable effects, calculated as percentage of baseline. This approach is advocated for providing guidance regarding strength of relationships and "dose equivalency" - not for statistical testing. It is well known that percentages (Turnage et al., 1987) lack sufficient statistical power and are generally to be avoided for analytic purposes. We use percent decrement here to provide a basis for comparison of effects across different treatments.

When these rational and experimentally controlled data from an alcohol study are used to "calibrate" or mark the other results, it would appear that the chemoradiotherapy treatments (Parth et al., 1988) exhibit the strongest effect although we also know (not shown) that this effect recovers when the subjects who survived the treatment were tested 12 months later. Note also that while scopolamine alone has a slight (and mildly significant) effect (Kennedy, Wood et al., 1988) when scopolamine is combined with amphetamine this effect is lessened. The altitude study (Banderet et al. 1988) shows a similar relation and even at the highest altitude obtained (23-25,000 feet - the approximate height of Mt. Everest) the effect is no stronger than we found with 2-3 drinks of alcohol (i.e., .05 - .10 BAC). Although the data are too sparse to conclude confidently, the pattern of the changes is illustrative of the conclusions which may be possible with a larger data base. For example, Grammatical Reasoning (symbol manipulation) appears to be most sensitive in all treatment conditions. Reaction Time, a response speed measure, also appears sensitive. Whether other treatments will show the same effect or not is problematic and awaits further study. We believe a completely filled matrix of tests X agents X dosages X mental factor would be extremely useful in setting limits for agents encountered in the military, space or the private sector.

In summary, APTS - a computerized test battery, has excellent metric qualities for repeated measures testing in a wide variety of applications. These range from a time course study for fitness-for-duty (e.g., fatigue, vigilance, or drug effects on performance) to sensitivity testing. Relations, metric properties, predictive validity and sensitivity studies provide a sound base for further data base development. Broader norms and more cohesive standardization efforts are needed.

Support for this project was provided under National Aeronautics and Space Administration Contract NAS9-17326 (COTR Dr. Frank Kutyna) and National Sciences Foundation Grant ISI-8521282 and by Essex IR&D funding.

TABLE 1. CHIEF STATISTICAL CRITERIA FOR APTS

STABLE OVER SESSIONS	
MEANS -	LEVEL, ASYMPTOTIC, ZERO RATE OF CHANGE IN SLOPE
VARIANCES -	CONSTANT OR CHANGING IN PROPORTION TO MEANS
INTRA-TRIAL CORRELATIONS -	CONSTANT (I.E., SYMMETRICAL)
RELIABLE	
RETEST CORRELATIONS	$r > .707$ FOR 3 MINUTES
FACTOR DIVERSITY	NON REDUNDANT TESTS
ECONOMICAL	CONSERVE TIME STABILIZE QUICKLY

TABLE 2. ESTIMATED TRIAL OF STABILITY, RELIABILITY, RELIABILITY-EFFICIENCY, AND FACTOR STRUCTURE OF THE SHORT AND MEDIUM LENGTH APTS*

Recommended Tests	Trial of <u>Stability</u>	<u>Task</u> <u>Defin</u>	Relia. <u>Effic.</u>	Factors			
				<u>MS</u>	<u>SMR</u>	<u>CPS</u>	<u>RSS</u>
Short Battery < 15 Min.							
NONPREFERRED HAND TAPPING	2	93	98	+ + +			+
4-CHOICE REACTION TIME	2	84	91				+ + +
CODE SUBSTITUTION	3	77	83		+ +		+ +
GRAMMATICAL REASONING	4	85	90	+ + +			+
PATTERN COMPARISON (Simult)	3	83	86		+ +		+ +
MANIKIN	3	86	92		+ +		+ +
TWO-HAND TAPPING	2	86	98	+ +			
Medium Battery < 20 Min.							
MATH PROCESSING	5	74	74	+	+	+ + +	
PATTERN COMPARISON (Succ.)	4	62	62		+		+ + +

Legend:

MS = Motor Speed

SMR = Symbolic Manipulation

CPS = Cognitive Processing Speed

RSS = Response Selection Speed

+++ = Loadings < 0.60

++ = Loadings 0.40 - 0.59

+ = Loadings 0.25 - 0.39

*All numbers reflect the average strength of relationships across eight separate studies.

TABLE 3. APPROXIMATE PERCENT DECREMENT ACROSS SIX SENSITIVITY STUDIES INVOLVING THE APTS

APTS Based Tests	Types of Intervention Levels of Treatment and Percent Functional Loss from a Placebo									
	Alcohol			Altitude		MS Drugs		Chemo	Drugs	
	.05*	.10	.15	15-20K	23-25K	Scop	Comb	Avg.	X	Y
Short Battery < 15 Min.										
NONPREFERRED HAND TAP				0	4	3	0	10	--	--
4-CHOICE REACTION TIME	5	10	23	--	--	--	--	12	11	10
CODE SUBSTITUTION	7	9	25	--	--	--	--	37	3	2
GRAMMATICAL REASONING	9	9	20	4	22	9	7	39	17	7
PATTERN COMP. (SUC)	0	4	10	1	12	2	3	17	2	3
MANIKIN	6	9	14			2	0	15	1	0
TWO-HAND TAP	--	--	--	0	4	--	--	4	--	--
Medium Battery < 20 Min.										
MATH PROCESSING	6	4	10	--	--	--	--	--	--	4
PATTERN COMPARISON	--	--	--	--	--	--	--	--	--	--

Legend:

Drugs = Antihistimine

X = Drug X

Y = Drug Y

Halon = Fire extinguishing agent

MS = Motion Sickness Drugs

Chemo = Chemoradiotherapy

Scop = Transdermal Scopolamine

Avg = Average decrement across the evaluation

REFERENCES

- AERA, APA, NCME, (1985). Standards for educational and psychological testing. American Psychological Association, Inc.: Washington, D.C.
- Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks Cole.
- Banderet, L. E., Shukitt, B. L., Crohn, E. A., Kennedy, R. S., Smith, M. G., Houston, C. S., & Bittner, A. C. (1987). Cognitive performance and subjective responses during prolonged ascent to 7600 m (25,000 ft) simulated altitude. Manuscript submitted for publication.
- Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1986). Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 114 measures. Perceptual and Motor Skills, 63, 683-708.
- Dunlap, W.P., Jones, M.B., & Bittner, A.C., Jr. (1983). Average correlations vs. correlated averages. Bulletin of the Psychonomic Society, 21, 213-216.
- Dunlap, W.P., Jones, M.B., Kemery, E.R., & Kennedy, R.S. (1986, November). Optimizing a test battery by varying subtest times. Proceedings of the 28th Annual Conference of the Military Testing Association (pp. 225-230). Mystic, CT: U.S. Coast Guard Academy.
- Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.
- Jeanneret, P. R. (1988). Position requirements for space station personnel and linkages to portable microcomputer performance assessment (NASA TR- 172038). Washington, DC: National Aeronautics and Space Administration.
- Kennedy, R. S., Baltzley, D. R., Dunlap, W. P. Wilkes, R. L., & Kuntz, L. A. (1988). Microcomputer-based repeated measures tests: Metric properties and predictive validities. Orlando, FL: Essex Corporation.
- Kennedy, R. S., Baltzley, D. R., Osteen, M. K., & Turnage, J. J. (1988, October). A differential approach to microcomputer test battery development and implementation. Proceedings of the 32nd Annual Meeting of the Human Factors Society (pp. 838-842), Santa Monica, CA: Human Factors.
- Kennedy, R. S., & Bittner, A. C., Jr. (1977). The development of a Navy Performance Evaluation Test for Environmental Research (PETER): In L. T. Pope & D. Meister (Eds.), Productivity enhancement: Personnel performance assessment in Navy systems (pp. 393-408). Naval Personnel Research and Development Center, San Diego, CA. (NTIS No. AD A056047)
- Kennedy, R. S., Dunlap, W. P., Jones, M. B., Lane, N. E., & Wilkes, R. L. (1985). Portable human assessment battery: Stability, reliability, factor structure and correlation with tests of intelligence. Washington, DC: National Science Foundation. (NTIS PB88-116645/A03)
- Kennedy, R. S., Jones, M. B., Baltzley, D. R., & Turnage, J. J. (1988). Factor and regression analysis of a microcomputer-based cognitive test battery. Orlando, FL: Essex Corporation.
- Kennedy, R.S., Lane, N.E., & Kuntz, L.A. (1987, August). Surrogate measures: A proposed alternative in human factors assessment of operation measures of performance. Proceedings of the 1st Annual Workshop on Space Operations, Automation, & Robotics (pp.551-558). Houston, TX: Lyndon B. Johnson Space Center.
- Kennedy, R. S., Wood, C. S., Baltzley, D. R., Odenheimer, R. S., & Dunlap, W. P. (1988). Effects of scopolamine and amphetamine on microcomputer performance tests. Orlando, FL: Essex Corporation.
- Lane, N. E., & Kennedy, R. S. (Eds.). (1988, May). Users manual for the Essex Automated Performance Test System (APTS) (Tech. Rep. No. EOTR 88-5). Orlando, FL: Essex Corporation.
- Parth, P., Lane, N. E., Dunlap, W. P., Chapman, R., Kennedy, R. S., & Ord, J. M. (1988). Cognitive deficits resulting from chemoradiotherapy in bone marrow transplant patients. Orlando, FL: Essex Corporation.
- Turnage, J. J., Kennedy, R. S., & Osteen, M. K. (1987). Repeated measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations. Orlando, FL: Essex Corporation.
- Winer, B. J. (1971). Statistical principles in experimental design (pp. 515-516). New York: McGraw-Hill.

A Biodata Instrument for Civil Service Examinations: Initial Validation

Jay A. Gandy, Alice N. Outerbridge, and James C. Sharf
U. S. Office of Personnel Management

The U. S. Office of Personnel Management has a continuing interest in developing applicant selection procedures for Federal civilian employment which have high validity and low adverse impact on minorities. The recent literature (e.g., Reilly & Chao, 1982; Owens, 1976) offers favorable indications that empirically scored biodata (life history, autobiographical, or background) information may be a useful component in the total assessment process.

OPM undertook a study in the Fall of 1987 to evaluate the use of biodata for entry-level (GS 5/7) professional and administrative (PAC) occupations. The PAC category includes over 100 occupations, such as procurement specialist, computer specialist, bank examiner, and personnel specialist. Trainees for PAC positions are frequently hired directly out of college, although a college degree is not a requirement. This paper discusses the development and initial validation of an empirically scored biodata questionnaire for potential use in selection for these occupations.

The Civil Service Commission (the predecessor organization to OPM) previously considered initiating research on biodata (McKillip & Clark, 1974) but concluded that the difficulties were too great and the investment overly risky. It was assumed at the time that biodata instruments had weak job relatedness and tended to invade personal privacy; also that validities tended to be occupationally and situationally specific and tended to deteriorate over time.

We recently took a fresh look at the issue. Recent developments reflected in the literature (e.g., Mumford and Owens, 1987) gave reason for optimism concerning the potential usefulness of biodata, and we anticipated that a careful development strategy might reduce potential problems with content acceptability. Additionally, there has been increased concern from many quarters that traditional civil service examining has failed to adequately assess "the whole person" with respect to characteristics important for job success. There is an urgent need for valid and practical tools to measure non-cognitive characteristics in addition to the abilities measured by written tests. This project is an effort in that direction.

Method

Instrument Development

Development of the instrument began with the taxonomies of past behavior items compiled by England (1971) and Glennon, Albright, and Owens (undated). We reviewed the biodata item dimensions against five decision rules we developed to evaluate acceptability in relation to privacy and related concerns. To be acceptable, items had: (1) to deal with events which were under control of the individual; (2) to

Opinions expressed in this paper are the authors' and do not necessarily represent the official policy of the U.S. Office of Personnel Management.

have some apparent relevance to job performance; (3) to be verifiable (in principle); (4) to be unlikely to be perceived as invading personal privacy; and (5) to avoid stereotyping by race, sex, or national origin.

The prescreening process eliminated most dimensions of background experience from further consideration. Item content areas remaining related primarily to school and educational experience, work history, skills, and interpersonal relations. Thus, the research questionnaire was limited to biodata items in these areas.

The research questionnaire contained 148 multiple choice items, each with five response alternatives. Respondents were instructed to make a single choice among the five alternatives. The great majority of item responses formed a continuum; and most items contained an escape option whereby the respondent could indicate that none of the options applied.

Criterion Form

The supervisory performance appraisal form was developed to provide job performance criteria for developing the scoring key for the biodata items. One part of the performance appraisal form consisted of a slightly modified version of the Descriptive Rating Scale (DRS) which has been used for many years by the U. S. Employment Service in hundreds of General Aptitude Test Battery (GATB) validation studies and which served as the basis for GATB validity generalization conclusions. The first part consisted of five work dimensions which are generally applicable to virtually all jobs -- hence the designation "generic" ratings -- plus a summary performance rating. These "generic" ratings included "quantity," "quality," and "accuracy" of work, "job knowledge," and "efficiency."

The second part of the appraisal form provides for ratings on ten abilities which were derived from earlier job analyses conducted in conjunction with the development of the OPM Professional and Administrative Career Examination (PACE). Supervisors were asked to indicate whether the ability was important or not for the position being rated. Ability ratings were made on a five-point scale ranging from "well below average" to "well above average" in reference to "all others you have known at this employee's grade."

It was hypothesized that the "generic" ratings of work performance would yield larger validity coefficients in comparison to the ability-based ratings. Considerations which went into this prediction included the following:

- The "generic" ratings were judged to be comprehensive in their applicability to different occupations based on the U.S. Employment Service's research evidence. In contrast, the ability ratings were not expected to be equally applicable to all jobs. Although the abilities were expected to be important, based on job analyses of PAC occupations, they were not expected to cover all aspects of work performance.

- The "generic" rating form describes dimensions of work performance in terms familiar to and commonly used by supervisors in describing and evaluating work performance. The ability descriptions are less familiar and probably require more careful reading and consideration.

- The "generic" descriptions require direct evaluations of work performance. The ability descriptions pose the more difficult

task of inferring relative levels on the abilities from observations of work behavior.

The above hypothesis put to empirical test the widespread belief that criterion measures based on job-specific abilities are needed both to enhance correlations with predictors and job relatedness defensibility.

Procedure

Two separate research packages were developed -- one for the employee and one for that employee's supervisor. Employee participants were provided a package of materials including instructions, questionnaire, answer sheet for optical scanning, and return envelope. Supervisors were provided with an instruction sheet explaining the performance appraisal criteria, an optically scannable answer sheet to record their appraisals, and a return envelope. Both employees and supervisors were assured of confidentiality, and all questionnaires and completed response sheets were mailed directly to OPM's processing center in Macon in the provided preaddressed envelopes.

Research Sample

The research sample was identified through OPM's Central Personnel Data File (CPDF). Research packages were sent to 15,300 employees and to each employee's supervisor. Most of this sample (13,000) consisted of all external hires at the GS 5/7 level for PAC occupations through all appointment authorities in all agencies for calendar years 1983 - 1986 who were still on board in PAC occupations as of June 1987. The remaining 2,300 was a random sample from 42,000 employees who entered PAC occupations through inservice appointments during the same period and who were still in PAC occupations as of June 1987. Local personnel offices were requested to identify the supervisor of each sample member and to deliver the performance appraisal materials.

Completed forms were returned by over 8,400 employees and 9,000 supervisors. Supervisor-employee matches were obtained for 6,300 cases, resulting in a 42% useable return rate. This sample was found to be representative of the target population with respect to relevant variables, including gender, race and national origin, occupations, and agencies.

Item Keying and Scoring

The scoring key for weighting item alternatives was developed empirically. Of various methods of keying biodata items (empirical keying, rational keying, factorial keying, and subgrouping), empirical keying has been found to be equal or superior to other methods when the purpose of the instrument is personnel selection (Mumford & Owens, 1987). Point-biserial correlations were computed between each item response and the criteria. A significance level of .01 was applied to item keying for split-half analyses (discussed below), and the .001 level was used for total group analyses. Additionally, rational decision rules were developed to supplement the statistical decision rule. Escape options, for example, were given a neutral weight. Unit weights of -1, 0, +1 were used in keying alternatives. Keying was

carried out independently by two psychologists. Very few keying differences occurred and these were discussed and resolved.

Validation Method

A double cross-validation design was used to determine the stability of predictor-criterion relationships. This procedure entailed the following steps: (1) splitting the total sample into two random halves; (2) developing a scoring key on each half; (3) applying each key developed on one half to the other half; (4) correlating the scores so derived with the job performance measure; and (5) evaluating the degree of "shrinkage," resulting from the independently derived keys. The final key was developed on the total sample, since the larger sample size would be expected to yield the most stable key for operational use. This process was repeated for different criterion combinations.

Subgroup Analyses

Separate analyses were conducted by gender and for race/national origin subgroups which had N's larger than 200. Gulliksen-Wilkes (1950) regression analyses were used to test for differences in standard errors, slopes, and intercepts.

Results

Score and Criterion Distributions

The biodata score distribution was approximately normal. Performance rating distributions were highly negatively skewed, but did have greater variance than operational (administrative) performance appraisals tend to have with about 10% of the employees rated below fully satisfactory.

Double Cross-Validation

One set of analyses used the average rating across all performance appraisal ratings as the criterion. This criterion combined all "generic" criteria plus those ability criteria rated as important for successful job performance. With the key based on half A of the sample, the correlation was .32, shrinking to .30 when this key was applied to half B of the sample. When the key was based on half B, the correlation was .31 with the half B sample, shrinking to .29 when the key was applied to the opposite half. The correlation of the key developed on the total sample was .30 with each half and also with the total sample.

A second set of analyses was carried out using the average rating across "generic" criteria alone. As hypothesized, this criterion measure yielded higher correlations ($p < .01$, non-independent r 's), although the differences are not large. With the key developed on half A, the correlation with the performance ratings of half A was .34, and the correlation on half B was .33. With the key developed on half B, no shrinkage was found: The correlation increased to .34 when the half B key was applied to half A. The key developed on the total sample correlated .33 with each half and also with the total sample. This key resulted in a total of 84 weighted items.

Subgroup Analyses

Subgroup analyses were conducted against both of the criterion measures discussed above, with similar results. Females tended to score about one-quarter of a standard deviation higher than males. No unfairness with respect to gender was indicated, however, based on Gulliksen-Wilkes regression analyses.

Similar analyses likewise showed no unfairness for Blacks and Hispanics. Regression analyses indicated that the relationship between scores and job performance is similar for all groups, although intercepts were significantly different. Results indicated moderate "overprediction" for Blacks and, to a smaller degree, for Hispanics.

The mean biodata score (T-scores) of Blacks (47.25) was about a third of a standard deviation (.34 SD) below that of Whites (50.65); and the mean score of Hispanics (48.72) was less than one-fifth of a standard deviation (.19 SD) below that of Whites. The small differences in Black/White means -- although based on incumbents, not applicants -- was encouraging relative to the differences typically found with ability tests (which frequently result in mean score differences of one standard deviation or more). Even so, additional analyses were undertaken to determine whether the observed differences could be further reduced without substantial loss in validity. To do this, all items were closely reviewed by subgroup at the item response level. Items having the lowest correlations with the criterion for minorities, and which also were characterized by relatively higher minority response rates for low-weighted alternatives, were successively removed. The decision process also took item content into consideration so that biodata dimensions were maintained. The biodata dimension measuring "academic achievement" was the dimension most heavily affected by deletions. Ultimately, 20 items were removed, leaving 64 weighted items in the questionnaire.

The effort to reduce adverse impact without compromising validity was moderately successful. The differences between Whites and minority subgroups decreased, and overall validity dropped very slightly (.33 to .32). The White-Black effect size decreased from .34 SD to .28 SD; and the White-Hispanic effect size also fell (.19 SD to .09 SD). Male-female differences also narrowed from .25 SD to .11 SD.

Discussion

The results of the initial evaluation using a concurrent validation model are encouraging. In an operational setting, validity would be expected to be subject to both positive and negative influences. From positive standpoints, an applicant sample would likely be less restricted on variables or characteristics which contribute to validity. From negative standpoints, motivational factors would be different for applicants relative to incumbents, and this could serve to increase error variance.

A validity level of .30 would be highly useful in selection when viewed from the perspective of expectancies for highly successful performance. Presently, about 54% of the PAC employees are considered to be highly successful. Success rates for those who scored above the

mean on the biodata questionnaire are substantially higher, ranging from 54% to 76%. Conversely, failure rates (i.e. ratings of less than fully successful) are only 2% at the highest score levels on the biodata instrument and range to 25% at the lowest score levels. The base rate for failures is about 8%.

Current and planned research is focusing on four areas: (a) conducting validity generalization analyses across occupations and agencies, (b) studying what the instrument measures in terms of construct validity, (c) evaluating the stability of validity with applicants versus employees, and (d) determining the impact on minorities in the applicant population compared to pre-selected incumbents.

With respect to validity for applicants, we will anticipate and develop steps to deal with potential problems of faking and score inflation. We expect to carry out verification activities on samples of items and samples of applicants and, as importantly, to make applicants aware that responses may be subject to review and verification. The literature base is very weak in the area of validity of self-assessments in employment contexts; but available studies such as Mabe and West's (1982) meta-analysis of studies involving self-evaluation provide some indication that anticipation of verification is an important factor in the validity of self-assessments.

References

- England, G. W. (1971). Development and use of weighted application blanks (Rev. ed.). Minneapolis, MN: University of Minneapolis, Industrial Relations Center, No. 55.
- Glennon, J. R., Albright, L. E., and Owens, W. A. Catalog of Life History Items. (c. 1961). (Compiled for Scientific Affairs Committee of Division 14, American Psychological Association).
- Gulliksen, H. & Wilks, S. S. (1950). Regression tests for several samples. Psychometrika, 15, No. 2.
- Mabe, P. A. & West, S. G. (1982). Validity of self-evaluations of ability: A review and meta-analysis. Journal of Applied Psychology, 67, 280-296.
- McKillip, R. H. & Clark, C. L. (1974). Biographical data and job performance. Washington, DC: U. S. Civil Service Commission (now U. S. Office of Personnel Management).
- Owens, W. A. (1976). Background data. In M. D. Dunnette (Ed.), Handbook of industrial psychology. New York: Rand-McNally.
- Reilly, R. R. & Chao, G. T. (1982). Validity and fairness of some alternative employee selection procedures. Personnel Psychology, 35, 1-63.

Synthetic Validation Procedures for Identifying Selection Composites and Cut Scores¹

Jane M. Arabian, U.S. Army Research Institute
Jeffrey J. McHenry and Lauress L. Wise, American Institutes for Research

In Project A, the Army is validating ASVAB (Armed Services Vocational Aptitude Battery) and an experimental battery of predictor tests for 19 enlisted military occupational specialties (MOS). Results indicate that the ASVAB and several of the experimental predictors are highly valid predictors of job performance. The problem now confronting the Army is to determine how to take advantage of the Project A data to assign selection composites to all 265 MOS open to Army enlistees. This paper will describe how we are using synthetic validation procedures to help the Army solve this problem. In addition, we will describe performance standard setting research that also is being conducted as part of the project. The purpose of this research is to develop procedures that will enable the linking of job performance standards to selection test cut scores.

During the Project A concurrent validation data collection, a battery of job performance measures and experimental predictor tests were administered to first tour soldiers in 19 military occupational specialties (MOS) selected to be representative of the 265 MOS open to first-tour soldiers. The performance measures included hands-on tests, paper-and-pencil knowledge tests, Army-wide and job-specific performance ratings, and self-reports of information contained in personnel files (e.g., letters of commendation, Articles 15). The predictors included six paper-and-pencil spatial tests, ten computerized psychomotor and perceptual tests, an interest inventory, and a temperament/biodata questionnaire.

Analyses of the Project A criterion measures indicated that there are five factors underlying the job performance of first-tour soldiers: Core Technical Proficiency, General Soldiering Proficiency, Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing (Campbell, McHenry and Wise, 1987). Validity analyses of the concurrent validation proved that a single selection composite could not maximize the validity of performance predictions in all Army MOS (Wise, Campbell, and Peterson, 1987).

The next challenge, therefore, was to determine how to take advantage of the Project A data to assign selection composites to jobs -- including not only the 19 MOS studied for Project A, but also the remaining 245 MOS not included in Project A plus any new MOS that might be created in the future. In addition, procedures were needed for setting cut scores on these selection composites that were tied to MOS performance requirements. The Synthetic Validation Project was designed to address these two problems.

The synthetic validity approach was first introduced by Lawshe (1952) as an alternative to the situational validity approach, which requires separate validity analyses for each job in the organization, and the generalized validity approach, which assumes the validity of a single test across a range of similar jobs. Gunion (1976) provides a review of several approaches to conducting synthetic validation. The approach that we adopted involves the

following steps:

- o Identifying a set of job components that are applicable across a range of jobs
- o Identifying selection composites for predicting performance on each job component (e.g., using traditional criterion-related validity procedures to identify valid potential predictors for each component)
- o Developing predictor composites for each job by combining the selection composites for each job component, weighting the composite for each component in accordance with the component's criticality to the job.

A review of alternative types of job component models identified four basic approaches to job/task classification (cf. e.g., Fleishman and Quaintance, 1984): behavior description (e.g., handling objects), behavior requirements (e.g., decision making), ability requirements (e.g., finger dexterity), and task characteristics (e.g., fires main gun). We decided to combine behavior requirements and ability requirements approaches and to proceed with three approaches.

The first approach is a Task Model. An initial list of task categories was developed during Project A from duty area descriptions for the 110 enlisted jobs with the largest number of incumbents. The list was refined on the basis of a series of workshops with job incumbents in a variety of MOS. The current list includes 96 categories. Example categories include "Troubleshoot mechanical systems," "Operate radar," "Prepare technical forms and documents," and "Fire individual weapons." We are asking NCOs and officers to rate the importance of each task category for both Core Technical and General Soldiering Proficiency and for Overall Performance in their MOS so that we can weight each component appropriately when we derive selection composites for these two performance factors.

The second approach is a Behavior Model. The components in this model are general job behaviors that underlie performance on several job tasks; that is, the behaviors are not task specific. The taxonomy, which presently includes 53 job behaviors, is based on seven general behavior categories: interpersonal behaviors, speaking behaviors, writing behaviors, cognitive behaviors (information processing), cognitive problems solving and troubleshooting operating equipment and physical activities. Specific examples of behaviors from the taxonomy include "Coach peers," "Relay oral instructions," "Judge path of moving objects," and "Walk long distances." Again, we are asking NCOs and officers to rate the importance of each behavior for both Core Technical and General Soldiering Proficiency and for Overall Performance in their MOS.

The third approach we developed is an Attribute Model. In this approach, the components are job requirements described in terms of mental and physical abilities, interests, traits, and other individual difference dimensions. The model currently includes 31 attributes. Examples include "Verbal ability," "Eye-limb coordination," "Physical endurance," "Conscientiousness," and "Interest in technical activities." With this model there is a one-to-one correspondence between the predictors and the job components; the predictors

the measures of the job components. Thus, the model eliminates the need to establish links between predictors and job components. We are asking NCOs and officers to rate the validity of each attribute for predicting both Core Technical and General Soldiering Proficiency in their MOS.

We have begun to compare the performance of these three component models, using the following evaluation criteria:

- o reliability of component judgments
- o acceptability to subject matter experts (SMEs)/users
- o validity of synthetic selection composites.

Presently we are in the midst of a three-phase, iterative data collection that eventually will include 25 MOS. For Phase I, we conducted workshops with NCOs and officers from three MOS: Infantryman, Vehicle and Generator Mechanic, and Administrative Specialist. These soldiers provided the importance and validity judgments described above. We also asked the soldiers to evaluate how well the three models covered the tasks, activities, and attribute requirements of their jobs.

The single-rater reliability of the Core Technical importance and validity judgments is shown in Table 1. There were few large differences in reliability across rating type or MOS. These data indicate that we could achieve a reliability of .90 for mean importance or validity ratings with 10-15 SMEs per MOS. We found that the ratings made by officers tended to be slightly higher than the ratings provided by NCOs, but the pattern of ratings was very similar across the two groups and across posts. Results were comparable for General Soldiering Proficiency.

Table 1

Single-Rater Reliability of Job Component Judgments for Army Judges

Job Component Model	MOS		
	Infantryman (11B)	Vehicle Mechanic (63B)	Administrative Specialist (71L)
Task Categories*	.52	.36	.40
Job Behaviors*	.36	.23	.43
Attributes	.31	.34	.45

*Based on Core Technical importance rating.

**Based on Core Technical validity rating.

We also compared ratings across MOS. We computed mean job description "profiles" for the task, behavior, and validity ratings. We then correlated these profiles across MOS (see Table 2). Most of the profile correlations were in the .80s for General Soldiering Proficiency. This indicates that soldiers in all three MOS agree about the tasks and behaviors that comprise the common soldier portion of their job and the attribute requirements underlying successful performance of General Soldiering responsibilities. The

correlations were considerably lower for Core Technical Proficiency. This indicates that our SMEs were able to use the three models to identify the unique tasks, behaviors, and attribute requirements of their MOS.

Table 2

Correlation of Mean Job Description Profiles Across MOS

Performance Area	Type of Model	MOS		
		11B and 63B	11B and 71L	63B and 71L
Core Technical Proficiency	Tasks	.52	.19	.39
	Behaviors	.26	.11	-.06
	Attributes	.52	.49	.35
General Soldiering Proficiency	Tasks	.87	.84	.87
	Behaviors	.86	.77	.66
	Attributes	.87	.87	.78

During the workshops, we asked SMEs to identify any components that were missing from the taxonomies and to rate how well each taxonomy covered their jobs. The SMEs identified very few missing components. They indicated that the taxonomies covered virtually all important job responsibilities and requirements.

We now are in the process of comparing the validity judgments against other sources of validity data. From Project A, we have empirical validity estimates for Core Technical and General Soldiering Proficiency for these MOS. We will be able to compare these with the direct validity estimates supplied by our SMEs. This will enable us to "validate" the Attribute Model.

It will be somewhat more difficult to assess the validity of the Task and Behavior Models. Our SMEs did not supply validity estimates for each task and behavior in these models. Moreover, although we have empirical validity estimates for some attribute X task and attribute X behavior combinations from Project A, there are many missing combinations. Consequently, to test these models, we are collecting validity judgments for each combination from a group of approximately 50 psychologists with extensive test validation experience. We will use these estimates, plus the importance weights provided by the SMEs, to derive selection composites for Core Technical and General Soldiering Proficiency. We will compare these composites to the empirical selection composites from Project A to assess the validity of our synthetic validation procedures.

As noted previously, the second objective of the Synthetic Validation Project is to provide a rationale for establishing minimum qualifying scores on the selection composites. Our general approach to this problem is to identify performance standards for each MOS, then to link these standards to scores on the selection composite.

We are establishing multiple performance standards for each composite. These standards correspond to outstanding performance, acceptable performance, marginal performance, and unacceptable performance. We believe that multiple standards may prove especially useful in setting and evaluating training objectives and in estimating job performance requirements during the design of new systems.

We recently completed a literature review on standard setting that covered procedures for setting standards. We learned that virtually no research has been conducted on procedures or techniques for setting job performance standards; virtually all standard setting research has been conducted in educational or licensing/certification settings. Therefore, we are exploring three alternative procedures for setting job performance standards.

The first technique is the soldier-based technique. We asked our SMEs to indicate the percent of outstanding, acceptable, marginal, and unacceptable soldiers in their MOS for each of the five Project A job performance factors. We will be able to use Project A data to link these percentages to scores on performance measures and to identify cut scores for each of the four performance levels.

The second technique is the critical incident technique. Critical incidents and retranslation effectiveness ratings were collected from NCOs and officers during the development of the Project A Army-wide and MOS-specific rating scales. As noted above, these scales were used to help assess three of the five performance factors (Effort and Leadership, Personal Discipline, and Physical Fitness and Military Bearing). For the Phase I Synthetic Validation data collection, we sampled critical incidents from each of these performance factors and asked SMEs to indicate whether the incident was an example of outstanding, acceptable, marginal, or unacceptable performance. We will be able to link these judgments to effectiveness ratings and to scores on the three performance factors. From these, we will be able to establish cut scores on the three selection composites.

The third technique is the task-based technique. We divided the Project A examinees within each MOS into deciles on the basis of their Core Technical Proficiency score. For each decile, we computed the mean score on each of the Project A hands-on tests. We presented these 10 sets of hands-on test scores to the SMEs at the Phase I data collection workshops and asked them to rate each set outstanding, acceptable, marginal or unacceptable. We then followed the same procedure to obtain judgments for General Soldiering Proficiency. We will be able to use these judgments to identify performance standards for Core Technical and General Soldiering Proficiency that can, in turn, be linked to scores on selection composites.

We are investigating a number of data collection and measurement issues as part of our standard setting research. For example, during the data collection workshops, some SMEs receive normative data from Project A to aid them in making their critical incident and task-based judgments, while others do not. Similarly, some SMEs complete a practice exercise and discuss their practice judgments before completing the judgment task, while others receive no practice. In addition, we are assessing the job experience and job knowledge

of our SMEs to help us establish "SME qualifications" for the judgment task. The goal of this research is to identify the data collection procedures and judges that provide the most reliable standard judgments. The measurement issues that we will be investigating include the best method for combining standards for different job components into an overall performance standards (e.g., multiple hurdles model vs. compensatory model) and procedures for linking selection test scores to overall performance standards.

It is expected that the synthetic validation procedures developed during this project will provide the Army with a cost-effective means of identifying selection composites. Our Project A experience indicates that it costs approximately \$500,00 per MOS to conduct a thorough validation study. If our synthetic validation procedures succeed, we estimate that it will cost only about \$15,000 per MOS to identify "synthetic" selection composites.

The standard setting procedures should also lead to cost-effective personnel readiness. Standards that are too high will lead to excessive recruiting costs, while standards that are set too low will lead to excessive training costs and attrition. These costs can be avoided if performance standards and selection cut scores are set at an appropriate level.

References

- Campbell, J. C., McHenry, J. J., & Wise, L. L. (1987, April). Analysis of criterion measures: The modeling of performance. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.
- Fleishman, E. A., & Quaintance, M. K. (1984). Taxonomies of human performance: The description of human tasks. Orlando: Academic Press, Inc.
- Guion, R. M. (1976). Recruiting, selection, and job placement. In M.D. Dunnette (Ed.), Handbook of industrial and organizational psychology (pp. 777-828). Chicago: Rand-McNally.
- Lawshe, C. H. (1952). Employee selection. Personnel Psychology, 5, 31-34.
- Wise, L. L., Campbell, J. C., & Peterson, N. G. (1987, April). Identifying optimal predictor composites and testing for generalizability across jobs and performance constructs. Paper presented at the Second Annual Conference of the Society for Industrial and Organizational Psychology, Atlanta, GA.

¹The views expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Institute or the Department of the Army.

PSYCHOMETRIC PROPERTIES OF THREE ADDITION TASKS
WITH DIFFERENT RESPONSE REQUIREMENTS

Louis E. Banderet, PhD, Barbara L. Shukitt, B.A.,
SSG Michael A. Walthers
U.S. Army Research Institute of Environmental Medicine
Natick, MA 01760-5007 USA

Robert S. Kennedy, PhD
Essex Corporation
1040 Woodcock Road, Orlando, FL 32803 USA

Alvah C. Bittner Jr., PhD
Analytics, Inc.
2500 Maryland Rd., Willow Grove, PA 19090 USA

Gary G. Kay, PhD
Georgetown University Hospital
3800 Reservoir Road NW, Washington DC 20007-2197 USA

ABSTRACT

The psychometric properties of performance tasks are influenced by many variables, e.g., task demands, sensory modality tested, requirements for transfer of skills or new learning, and the subject's output response. Previously, we administered two Addition Tasks, where each problem required mentally summing three two-digit numbers. The format, spatial layout, and problem characteristics were similar; however, each task required a different output response (writing each answer on paper versus entering each with number keys on the top row of the computer keyboard). Although we gave subjects "touch typing" practice on the number keys, subjects responded 65% as fast on the automated task and its differential stabilities were less. To overcome the limitations of the computerized Task, we developed another Addition Task for computer administration that required only a dichotomous response ("Right" or "Wrong"). Problem format was similar to earlier versions, except each problem had a suggested sum and it was wrong on 50% of the problems. Subjects mentally calculated each sum, determined if each suggested sum was right or wrong, and entered their answer on the keyboard.

Recently, we evaluated the three tasks in a study which manipulated environmental stressors. The first computer task, requiring entry of the sum with the number keys on the computer, was psychometrically inferior to the other two. The second computer task was superior to even the paper and pencil task although its problem solving rate was 20% lower. These results demonstrate how relatively subtle differences in the response requirements of a task may dramatically influence its psychometric properties. Such findings suggest that traditional measures of neurological, cognitive, or perceptual status must be implemented cautiously on alternate media.

Technical advances in personal computers make them very useful for automated testing of questionnaires, personality inventories, and performance tasks. When traditional assessment instruments are adapted for administration by computer, changes in task characteristics or output responses are sometimes inevitable. For example, writing numbers on a paper and pencil task versus entering them with number keys on a task administered by computer. Hence, implementing tasks on the computer may change their psychometric properties. Testing tradition implies that the computerized instrument should be validated. Indeed, recent evaluations with systematic criteria suggest such caution is warranted (Bittner, Carter, Kennedy, Harbeson, & Krause, 1986; Bittner, Smith, Kennedy, Staley, & Harbeson, 1985; Farrell, 1983; Smith, Krause, Kennedy, Bittner, & Harbeson, 1983).

Recently, we adapted a paper and pencil Addition Task for administration on a portable computer. We found that it seemed insensitive to experimental effects and its response rate was 65% of that for the paper and pencil task. Therefore, we developed another Addition Task to overcome these shortcomings. The purpose of this investigation was to determine the psychometric properties of three Addition Tasks with different response requirements.

METHOD

Subjects---Twenty medical research volunteers from Fort Detrick, MD, and the Natick, MA, Research Development and Engineering Center were subjects. All were given physicals; no subjects had medical histories that would contraindicate altitude and cold exposure. Human subjects participated in these studies after giving their free and informed voluntary consent. Investigators adhered to AR 70-25 and USAMRDC Regulation 70-25 on Use of Volunteers in Research.

Psychometric Instruments---Three Addition Tasks were evaluated (see Figure 1). All tasks were timed and required summing three two-digit numbers to solve each problem; however, the output response differed for the three tasks. The administration medium also differed for some.

The P-SUM Addition Task was a traditional paper and pencil task, generated by computer (Carter & Sbisá, 1982), with 60 problems per page. A subject could change an answer anytime during the testing interval.

The C-SUM and C-RorW Addition Tasks were administered on a GRID Compass portable computer with a high contrast electro-luminescent display. The spatial arrangement of the 40 problems

per display for these tasks was similar. The C-SUM task required entering each three-digit sum with the number keys on the top row of the keyboard. The cursor's position below a problem prompted for the sum's digits in a specific order. The rightward digit of each sum was entered first, then the leftward, and finally the center digit. An arithmetic or entry error could not be changed after a digit was entered on the keyboard.

The C-RorW Addition Task presented a suggested sum for each problem which was wrong on 50% of the problems. Subjects mentally calculated each problem's sum and decided if each suggested sum differed from the correct sum. The task was programmed for deviations of +1 in the 100, 10, or 1-digit of the sums on 2.5, 45, or 2.5% of all problems, respectively. Subjects pressed the 'F' or 'G' key if the suggested sum was correct; they pressed the 'H' or 'J' key if it was wrong. Like the C-SUM task, a response could not be changed after a key was pressed.

Procedures---This study was part of a larger investigation to determine if an amino acid, tyrosine, prevents some of the adverse behavioral effects induced by environmental stressors. Specifically, on three occasions two groups of 10 subjects were exposed to 4700 m of simulated high altitude and 17°C for 7 h and investigated with a repeated-measures design. Conditions were blinded as to administration of placebo and tyrosine (85 mg/kg and 170 mg/kg). Two other occasions, subjects were given placebo and tested at 500 m and 22°C. Before the experiment, subjects practiced each task 15 times (4 min per administration). Subjects were trained to work quickly with less than an 8% error rate. Feedback was given during training to facilitate rapid acquisition of the tasks and enhance motivation.

During experimental testing, subjects also performed each task for 4 min. Physiological and biochemical measures and an extensive symptom, mood, and performance battery were administered 90-420 min after ascent to simulated high altitude. The Addition Tasks were administered from 320-350 min, immediately after the third blood sample. The tasks were always investigated in the same order: C-RorW, P-SUM, and C-SUM. Map Compass, another task that we developed, and Number Comparison (Carter & Sbisà, 1982) were administered before and after the P-SUM task, respectively, to separate the three Addition Tasks. Sea level performance values were the average of the two sessions at 500 m; 4700 m values were from the day when the placebo was given to each subject.

Performance on each task was specified as number of problems correct per minute. To discourage careless responding, each error was doubled in the calculation of this index. To determine task sensitivity, a z score was calculated for each task, reflecting both the magnitude and variability of measured altitude effects. To obtain another estimate of the relative sensitivities of the tasks, z scores (altitude effects) were calculated for each task for 10 samples randomly resampled from the original subject population. Each sample (N = 10) was drawn without replacement; however, the parent population (N = 20) was always reconstituted before another sample was selected.

RESULTS

Figure 2 shows scatterplots with correlation coefficients relating each Addition Task with each other task for the sea level + 22°C (500 m) and simulated high altitude + 17°C (4700 m) conditions. High altitude conditions produced significant impairments in performance on all three tasks ($p < .0001$) which were usually evident in the scatterplots as a leftward and downward displacement of data points for the 4700 m condition. All correlations in Figure 2 were statistically significant. The C-RorW task was consistently more highly correlated with the P-SUM task for both environmental conditions than the C-SUM task was. The two computerized versions of the Addition task were more similar to each other for both environmental conditions than either was to the P-SUM task. Performance rates differed on the three tasks; the C-SUM task produced the slowest rates.

Table I shows characteristics of the three Addition Tasks for the 500 and 4700 m conditions. Performance rates were greatest for the P-SUM task. At 500 m, rates on the C-RorW task were 80% as great; those on the C-SUM task were 60%.

P-SUM			
43	75	52	MEDIUM: PAPER & PENCIL
38	26	73	RESPONSE: WRITE SUM
<u>28</u>	<u>54</u>	<u>88</u>	
109	155		
C-SUM			
34	82	50	MEDIUM: COMPUTER
79	82	92	RESPONSE: ENTER SUM
<u>46</u>	<u>97</u>	<u>11</u>	
119			
C-RorW			
99	63	55	MEDIUM: COMPUTER
32	29	25	RESPONSE: ENTER R or W
<u>49</u>	<u>35</u>	<u>26</u>	
180	117	116	
R	W		

Figure 1. Sample problems, administration media, and output responses for three Addition Tasks.

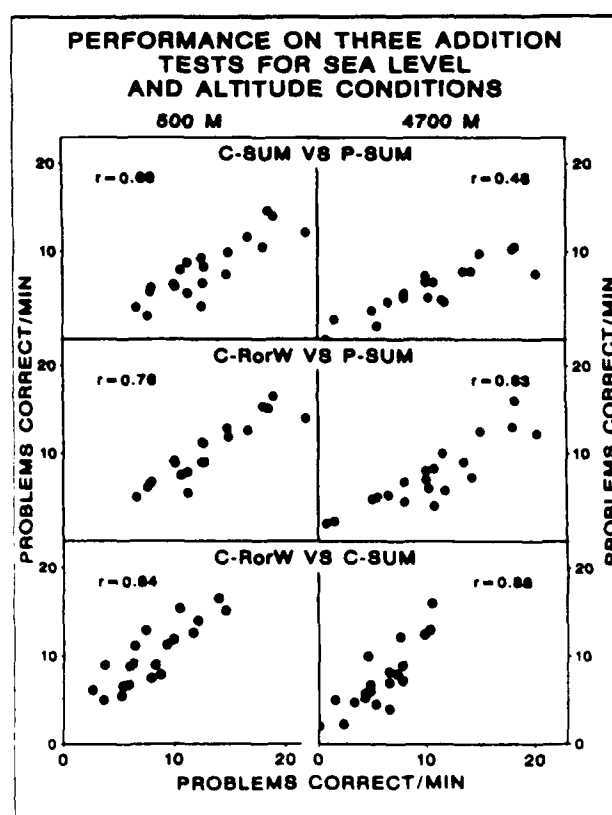


Figure 2. Scatterplots relating performance for each Addition Task with every other task at sea level + 22°C and simulated high altitude + 17°C conditions.

CHARACTERISTICS OF ADDITION TASKS AT 500 AND 4700 M

CRITERION	STATISTIC	P-SUM	C-SUM	C-RorW
500 M				
BASELINE RATES (CORRECT/MIN)	MEAN	12.87	7.97	10.11
	SIGMA	4.23	3.38	3.49
MINUTES PRACTICE (4 MIN/ADMIN)	MEAN	≤20	≤40	≤20
TASK DEFINITION (ADMINS 5 & 6)	PEARSON r	>0.85	>0.72	>0.85
RELIABILITY (4700 VS 500 M)	PEARSON r	0.76	0.70	0.93
4700 M				
ALTITUDE EFFECT (Δ CORRECT/MIN)	MEAN	-2.39	-2.22	-2.63
	SIGMA	1.97	1.85	1.84
	z SCORE	-1.22	-1.20	-1.43

Table I. Psychometric characteristics of three Addition Tasks for sea level + 22°C and simulated high altitude + 17°C conditions.

Variability was also greatest for the P-SUM task; it was less for the two computer administered tasks. The C-SUM task required twice as much practice as the other tasks to achieve comparable task stability ($> .80$). Task reliability for the experimental and control administrations was greatest for the C-RorW task. Task definition (test-retest reliability) was greatest for the P-SUM and C-RorW tasks. Table I also shows the mean altitude effects and standard deviations for each task. The C-RorW task was the most sensitive task as inferred from z score magnitudes; the P-SUM and C-SUM tasks were less sensitive.

Table II shows z scores and their ranks (1 = most sensitive) from our resampling procedure. With 10 samples, P-SUM was most sensitive twice, C-SUM was once, and C-RorW was seven times.

DISCUSSION

Performances on three Addition Tasks with differing response requirements were compared under control and stressful environmental conditions. This study demonstrated that our first computer implementation of a useful paper and pencil task (P-SUM) resulted in an automated task (C-SUM) with lower response rates, reduced sensitivity to experimental effects, less task definition, and greater practice requirements. It is of interest that this Addition Task is currently implemented in varied performance batteries. We suspect that entering numbers on a computer keyboard was awkward and incompatible with rapid responding since few of our subjects could "touch type".

NORMALIZED ALTITUDE EFFECTS					
RANKS			z SCORES		
P-SUM	C-SUM	C-RorW	P-SUM	C-SUM	C-RorW
3	2	1	-1.12	-1.26	-1.64
3	2	1	-0.80	-0.96	-1.15
2	3	1	1.08	-0.87	-1.31
2.5	2.5	1	-1.50	-1.50	-1.56
2	3	1	-1.24	-0.85	-2.04
1	2	3	-1.75	-1.65	-1.54
3	2	1	-1.42	-1.73	-1.82
3	1	2	-1.17	-1.69	-1.57
2	3	1	-1.29	-1.20	-1.59
1	2	3	-1.55	-1.12	-1.08
2.10	1.10	7.10	PROPORTION WITH LARGEST EFFECT		

Table II. Sensitivities of three Addition Tasks determined with data resampled from the original subject population.

To overcome the limitations of the initial computerized Addition Task, we developed a second Addition Task (C-RorW) which required a dichotomous response. Although the problem solving rate of the second computerized task was less than expected, its sensitivity and other psychometric properties were superior to the paper and pencil task. We suspect our second computerized task was more sensitive because it was more abstract and complex, e.g., subjects determined the sum for each problem, compared it to the suggested sum, and decided which response key to press.

Implementation of an Addition Task with an appropriate subject output response produced a computerized task that was psychometrically superior to its paper and pencil counterpart. In contrast, choice of the traditional output response for our first computerized task yielded the task that was the most inferior. These data illustrate the importance of evaluating tasks when they are changed or adapted to alternative media. They also demonstrate how relatively subtle differences in the response requirements of a task may dramatically influence its psychometric properties. Such findings suggest that traditional measures of neurological, cognitive, or perceptual status must be implemented cautiously on alternative media.

REFERENCES

Bittner, A.C., Jr., Carter, R.C., Kennedy, R.S., Harbeson, M.M., & Krause, M. (1986). Performance evaluation tests for environmental research: Evaluation of 114 measures. Perceptual Motor Skills, 63, 683-708.

Bittner, A.C., Smith, M.G., Kennedy, R.S., Staley, C.F., & Harbeson, M.M. (1985). Automated portable test (APT) system: Overview and prospects. Behavior Research Methods, Instruments, & Computers, 17(2), 217-221.

Carter, R.C., & Sbisa, H. (1982). Human performance tests for repeated measurements: Alternate forms of eight tests by computer (Report No. NBDL8213003). New Orleans, LA: Naval Biodynamics Laboratory.

Farrell, A.D. (1983). When is a computerized assessment system ready for distribution? Some standards for evaluation. Computers in Psychiatry/Psychology, 5(1), 9-11.

Computerized Testing (CAT) in the German Federal Armed Forces (GFAF)

Wolfgang Wildgrube
Psychological Service of the Federal Armed Forces
Armed Forces Office, Bonn, FRG

1. Introduction

In the Psychological Service of the Federal Armed Forces each year a great number of aptitude tests (about 350,000) is conducted applying most various psychological test methods. Most of these tests refer to the entrance diagnosis for draftees, for volunteers as enlisted personnel or for officer candidates. In addition there is a great number of special examinations, e.g. for pilot candidates, air traffic control personnel, tank drivers, but also for candidates for the civil service, interpreter service and so on. Especially in view of the personnel resources which clearly decrease, a much more differentiated and specific aptitude diagnosis is required especially with regard to placement aspects.

2. Status of the development of CAT in the Federal Armed Forces

Since 1983 the Federal Armed Forces have tested the application of computers in psychological test diagnosis within the scope of the "Computerized Adaptive Testing" (CAT) project. The aim is to achieve that the computer as an aid for the psychologist carries out the test examination as effectively as possible. With such an individualized and thus "tailor-made" test battery via computer it is possible to consider specifically the individual candidates' ability as well as the institutional interests of the Federal Armed Forces in the subsequent placement assessment done by the psychologist.

Since the beginning of the project in 1983 with equipment of the first personal computer generation (processor Z80) experiences and data from more than 20,000 computerized tests have been collected. Various presentations (see Wildgrube: MTA-Meetings 1985 and 1987) have shown that computerized tests can be conducted with great numbers of candidates without any problems worth mentioning. In addition, various accompanying inquiries have demonstrated the high acceptance of that type of psychological testing.

As standard entrance test for the Federal Armed Forces the aptitude classification test (German abbreviation EVT) is used. Since January 1986, after a major revision, it consists of the following procedures:

Figure Matrices Test (FDT)	20 items/8 alternatives	18 min.
Word Relation Test (WBT)	20 items/5 alternatives	4 1/2 min.
Arithmetic Reasoning Test	20 items/input of the results	14 min.
Spelling/Orthographical Test (RST)	60 items/4 alternatives	12 1/2 min.
Mechanical Ability Test (MKT)	20 items/5 alternatives	13 min.
Electrotechnical Comprehension Test (EKT)	20 items/5 alternatives (pretest after RT: 8 items)	20 min. 6 min.
Reaction Test (RP)	64 items/6 alternatives input of the results	3.41 min.
Radio Test (FT)	150 items/3 alternatives input of the results	3.30 min.
Signal Test (SigT)	18 items/4 alternatives input of the results	3 min.
Doppler Test (DopT)	20 items/3 alternatives input of the results	4 min.

At present the six initial subtests are still performed conventionally with paper and pencil, followed by the four apparatusive tests. This fixed test sequence is applied in the present routine testing at the Military Subregion Recruiting Offices and the Volunteer Recruiting Stations. In the computerized version the sequence may now be modified by the proctor by, for example, interruption of the test sequence, *cancellation of subtests or premature stop of the test battery* based on the results achieved in the initial subtests. This aptitude classification battery (EVT) as the basic diagnostic instrument in the Federal Armed Forces which is also used in the computerized version has essentially been extended in the meantime, *inter alia*, by extensive test procedures for the Officer Applicants Test Center in Cologne. Within the scope of the current CAT development efforts these entire test materials are to be integrated into the computer in order to make more differentiated test strategies possible. Compared with the present state more than 2,000 multiple choice items will then be usable.

The developments during the 1980s and the evaluation with draftees and volunteer candidates have shown that standard hardware and software are not sufficient to meet the manifold requirements of computerized testing. Therefore a detailed list of requirements was developed by the Federal Armed Forces and submitted to the developing companies. In accordance with the rapid change in the field of personal computer technology in 1984 the conversion from 8 bit processors (Z80/CAT I) to a technology with 16 bit processors (CAT II) was prepared. These advanced configurations are now tested in daily operation in the Recruiting Centers in Munich and Hildesheim.

During the extensive applications in these two Recruiting Offices, however, it also became obvious that for a comprehensive application in the area of volunteer selection more extensive computer capacities will be necessary. As a consequence the requirement descriptions were extended by the Psychological Service of the Federal Armed Forces so that at present an advanced computer configuration can be developed to equip the Volunteer Recruiting Stations and the Officer Applicant Test Center (CAT III). At present extensive software development is being accomplished with the aim of having the new CAT equipment generation operationally ready in summer 1989 in order to then be able to test applicants individually and specifically for the Federal Armed Forces in computerized test procedures as a matter of routine and in a wide scope and to subsequently provide psychological advice.

After an extension of the CAT II configurations (16 bit processors - processor standard 8086/8088) presently employed the following enhancements will occur:

- . The new hardware is based on computers with processor 80286 and co-processor 80287 with a 1 Mbyte memory each. The various work stations are connected via a Local Area Network (LAN), central data maintenance takes place in a separate file server. These test installations are monitored from a central proctor station supplemented by a second work station for a proctor.
- . Each candidate work station is equipped with a color screen and a headphone. The keyboard was specifically developed to meet the requirements of the various psychological tests and also includes a motoric input device to perform motoric tests.
- . By means of respective hardware and software design it is ensured that data losses are avoided and that after a breakdown in any case the test can be completed within one workday. In addition, the data are stored on a streamer band at the end of a test day taking into account the conditions determined by the Federal Data Protection Commissioner.
- . At each location of a CAT installation the entire item pool presently employed by the Psychological Service of the Federal Armed Forces is available. Apart from the aptitude classification test battery and the aptitude test battery for officer applicants also all special procedure.. are accessible on the computer. Thus at any point in time it is possible to test each candidate with specially designed procedures or with individually tailored test batteries for special psychological questions.
- . An extensive program tool called item editor enables the user to modify existing test procedures and to include new subtests. The modifications refer to item texts and graphs as well as to standard values and correct answers so that existing test procedures or new tests, respectively, can be modified by the Federal Armed Forces at any point in time.
- . For the candidate the testing session begins with an extensive learning phase during which he exercises to operate the keyboard and may become familiar with the upcoming test items. That learning phase as well as the instructions for the individual subtests are supported by additional test instruction via headphone. By means of these measures it is achieved that the understanding of the items is considerably increased especially for weaker draftees and that consequently better results are achieved in the individual tests.
- . Adaptive test algorithms are being tested in co-operation with the University of Aachen. Most recent experiences - also of the University of Vienna - show, however, that practical applications of such adaptive algorithms do not yet exist. However, due to the item editor concept it will be easily possible to implement respective subtests and methodic models on the CAT systems and to apply them for routine testing.

3. Problems of time measuring

Computerized test systems make it possible for the first time to collect data of conducted tests at short notice and to further evaluate them on a mainframe computer. After various difficulties in the technical implementation (data transfer from PC via magnetic tape to the mainframe) extensive data bases are now available for further analysis. Apart from the classical evaluations such as item and distractor analyses, first results are described in the following which refer to the evaluation of the latency per item. Computerized testing for the first time provides the possibility to record and systematically evaluate the latency per person and item in addition to the evaluation by "correct"/"incorrect".

The following three tables contain evaluations regarding the response time (latency) for the subtests "Word Relation Test" (WBT), "Figure Matrices Test" (FDT) and "Arithmetic Reasoning Test" (RT) of the aptitude classification test battery per subtest and form with at least 600 subjects. The aim of an evaluation of these data should be to obtain more advanced parameters on a person beyond the sum score of a subtest (sum of correctly solved items) in order to round off the diagnostic process and to obtain additional/more comprehensive information on an individual. For that purpose, inter alia, also the response time per item or times related to one subtest, respectively, are to be used when respective analyses will now be able to provide informative results/parameters.

Separately for form 1 or 2 the tables contain the item difficulties and discrimination parameters according to classical test theory. These figures are followed the average value and the variance of the response time (measured in seconds) per item and the minimum and maximum times. In the last two columns the correlations between difficulty and latency were entered plus the number of persons as the computation basis for that correlation.

In general the results of the presented three subtests on the basis of an item analysis according to the classical test theory are satisfying. The reliabilities range from .83 to .90 (Cronbach-Alpha). The difficulties cover a broad spectrum of the measuring range, the discrimination parameters for most of the items are between .25 and .55 (see especially the "Arithmetic Reasoning Test" - with a guessing probability = 0). The descriptive evaluations of the time (measured in seconds) per item do not provide a homogeneous picture. The range between minimum and maximum is very large in the three tests. The variation of the medium values is relatively small, especially for the Word Relation Test, however, combined with high dispersion values.

Finally the correlations between difficulty and response time in all three tests provide only small and mostly negative values. The number of persons included in the correlations clearly decreases towards the end of each test (especially in the case of the Arithmetic Reasoning Test). Consequently it is true that there are weak relations between difficulty and latency of an item but there is a great amount of unclarified variance between the two indicators of an item.

Since an interpretation of these first results regarding the response time of multiple choice items is not considered possible some conceivable explanations are presented. Accordingly the incorrect/correct processing of performance items and their processing speed belong to different latent dimensions. Thus the performance range could be set against personality dimensions in their closer sense such as cognitive styles, search for success/avoidance of failures or processing strategies which then will be expressed by the latency (separate evaluations according to school education are planned).

These first simple evaluation results should initiate discussions and advanced studies in which this additional individual parameter of the solution time is to be investigated by means of data collections guided by hypotheses/theory. Within the scope of further tests on CAT in the Federal Armed Forces and the planned routine application of computers in conducting tests of volunteers we will continue to collect specific data regarding those questions. The aim is to round off the diagnostic judgement to fulfil the requirements of various questions and activities in the personnel sector in the German Federal Armed Forces.

WBT Form 1 N = 654 / Alpha = .832

	Items		Latency (in seconds)				correlation	N
	diffi- culty	part-whole correlation	mean	standard- deviation	mini- mum	maxi- mum		
1	.789	.439	14.47	10.34	3	139	-.107	616
2	.856	.495	10.75	9.49	3	108	-.333	619
3	.768	.556	13.39	9.46	4	95	-.290	619
4	.677	.404	16.45	9.52	3	69	-.242	617
5	.546	.480	20.47	12.29	3	84	-.249	612
6	.642	.437	17.45	10.87	3	72	-.091	608
7	.696	.566	13.42	8.61	3	99	-.125	606
8	.569	.528	15.03	7.50	4	57	-.237	597
9	.515	.425	17.42	9.92	3	80	-.072	587
10	.624	.367	15.28	10.43	3	74	-.280	579
11	.320	.365	17.84	10.06	3	63	.031	565
12	.466	.505	14.66	7.89	4	62	-.108	535
13	.246	.313	15.67	7.82	2	47	.085	500
14	.315	.454	11.98	6.36	3	63	-.141	477
15	.280	.463	16.43	7.88	5	53	-.036	406
16	.324	.329	12.02	5.92	3	38	-.044	389
17	.197	.285	17.43	8.21	5	51	-.073	299
18	.098	.286	14.34	6.98	3	39	.226	230
19	.051	.225	13.68	6.87	3	45	.055	185
20	.061	.048	12.77	6.50	4	38	-.031	134

FDT Form 2 N = 612 / Alpha = .845

	Items		Latency (in seconds)				correlation	N
	diffi- culty	part-whole correlation	mean	standard- deviation	mini- mum	maxi- mum		
1	.807	.556	18.44	15.07	3	113	-.293	547
2	.856	.583	14.19	11.79	3	105	-.227	567
3	.745	.548	29.64	19.93	5	159	-.257	569
4	.675	.568	22.57	18.26	5	199	-.132	571
5	.680	.422	27.07	18.96	6	183	.023	568
6	.712	.583	44.63	27.36	8	182	-.055	562
7	.618	.481	28.98	17.86	4	148	-.060	571
8	.672	.542	47.31	33.82	4	317	-.097	563
9	.637	.509	40.34	24.57	6	186	.042	568
10	.425	.432	51.14	36.96	7	288	-.271	542
11	.652	.500	18.48	12.78	4	99	.083	568
12	.606	.539	36.74	24.83	4	232	.045	564
13	.562	.575	37.97	24.11	3	199	.288	564
14	.440	.339	26.27	16.30	4	151	.385	563
15	.281	.329	82.04	50.34	4	317	.120	523
16	.121	.132	76.36	44.68	3	306	.070	505
17	.051	.007	78.28	45.86	4	309	.042	460
18	.116	.169	66.48	43.28	2	228	.072	437
19	.074	.136	58.57	36.35	2	209	.122	404
20	.093	.117	45.61	29.79	5	199	.203	412

RT Form 2 N = 612 / Alpha = .905

	Items		Latency (in seconds)				correlation	N
	diffi- culty	part-whole correlation	mean	standard- deviation	mini- mum	maxi- mum		
1	.812	.315	30.98	21.69	6	190	-.172	558
2	.758	.542	36.64	31.01	7	234	-.252	545
3	.690	.449	42.25	25.95	5	281	-.331	553
4	.719	.546	45.22	31.78	9	206	-.230	531
5	.680	.568	37.94	35.05	7	240	-.266	531
6	.479	.609	73.60	48.66	11	267	-.133	422
7	.577	.547	50.04	38.70	6	269	-.205	515
8	.392	.626	66.16	40.22	13	302	-.232	343
9	.345	.641	66.33	37.36	16	267	-.323	333
10	.444	.644	62.93	35.21	13	245	-.241	418
11	.454	.578	47.07	34.56	6	240	-.304	456
12	.376	.564	37.35	44.07	13	368	-.029	418
13	.423	.601	44.52	25.03	12	196	-.152	415
14	.175	.574	57.75	39.25	8	217	.142	204
15	.235	.602	46.48	27.49	12	160	-.079	209
16	.276	.612	41.01	28.16	6	188	-.173	275
17	.229	.554	50.92	28.94	12	197	-.128	175
18	.108	.465	49.90	34.41	11	244	.181	175
19	.056	.400	40.86	28.04	8	165	.221	115
20	.034	.280	70.65	48.14	13	223	.145	60

**FUNCTIONALITY AND ARCHITECTURE OF THE GFAF
COMPUTERIZED ADAPTIVE TESTING SYSTEM**

Dr. Michael Habon
Instructional and Diagnostic Systems
Defense Division
DORNIER, FRG

Overview

The Instructional and Diagnostic Systems department of the DORNIER Defense Division is now developing the first full-scale implementation of a computerized adaption testing (CAT) system by order of the German Ministry of Defense. The CAT system is planned to be operational by July 1989 in the Officer Selection Center in Cologne and in the four Volunteer Selection Centers of the German Federal Armed Forces (GFAF) in Wilhelmshaven, Hanover, Düsseldorf, and Munich.

The functional requirements for the system are presented and its implementation is briefly discussed.

Functional requirements

The CAT system has to meet the following basic requirements: it shall be capable of being integrated without difficulties into the existing flow processes of the Volunteer Selection Centers, Officer Selection Centers, and the Conscript Examination and Placement Centers of the GFAF on the one hand and offer an extremely flexible environment for the administration of all test methods currently in use or under development.

To assist the processes based on work sharing in the various agencies in an effective manner the following conditions, outlined briefly in this report, had to be taken into account in the design decisions for the hard- and software of the CAT system.

1. It shall be possible to put in personnel data during the test session or at any other time at a special workstation via a keyboard or data medium. Personnel data and all test data for a person up to the item response level have to be collected and stored on tape. The password protected access to all data must be possible at any time by means of the "personnel-id-number". Different outputs have to be tailored to the requirements in the recruiting centers at subtest, person, or item level. Statistical evaluations of the data and text processing shall be feasible at this data input station.
2. The proctor's station shall offer the possibility of initializing, monitoring, interrupting, and terminating the test sessions centrally. The proctor shall have the possibility of administering standard test batteries or individual compilations of subtests.

Decision matrices, which have to be completed with subtest scores or composites, shall facilitate decisions during the session on the test breakoff or additional tests. During the session, the proctor shall continuously obtain information of the tests in hand, scores, test times, number of unsuccessful trials in the case of example tasks etc. In addition, he shall have the possibility of excluding individual subtests from a standard sequence, of making breaks, or stop the session at any time.

3. The testee at his workstation shall principally be in a position to work throughout the session without any external help. This means that instructions read to a group are presented individually via headsets, that legibility and identifiability of the items on the monitor is at least as good as with a printed copy, and that the input medium for the testee's responses can be handled without much practice. Moreover, the input medium shall be flexible, i.e. suitable for all kinds of current and future tests, for example for reaction, signal, doppler, and psychomotoric tests used in the Psychological Service. Adaptive administration of appropriately calibrated item pools must be possible.

In the case of system disturbances (power cutoff, hardware defects etc.) no data shall be lost, and the testee shall be in a position to continue the session on the same or another station with the last item displayed before interruption.

4. Users without programming knowledge in the Psychological Service shall be in a position to set up items, tests, and batteries or to modify existing procedures at a special workstation.

CAT system architecture

The analysis of the CAT system requirements led to the following hard- and software design decisions:

The CAT system is a network of AT-compatible personal computers. Within the network, a PC with two 40MB hard discs takes over the function of a dedicated server. The network, operated by a commercial operating system is fault tolerant, protecting against the loss of testee data since the latter are held on mirrored hard discs.

If the power supply is interrupted a self-contained power supply ensures the test operation for at least 10 minutes. During this period of time, the system can be set in a defined condition and switched off.

The personnel data from the data entry and data interpretation station are merged in the server with all the testee's test data according to the control criterion "personnel-id-number" and stored for a medium period of time until they are transmitted to a streamer tape for long-term data storage.

From the proctor station, a session is initialized at a certain testee station (both are requesters in the network) by writing a configuration file into this testee station's subdirectory on the server. This configuration file, provided with parameters by the proctor via a menu, includes a precise description of the session to be run on the testee station. The testee station asks for the file and automatically configures a session from parts of exercises, instructions, example tasks, and test items. To be able to follow this proceeding all text and graphic items of the Psychological Service (currently about 4000) and the flow process control of all tests (about 60) must be available at the testee station.

After each item response during the session, diagnostically relevant data are entered into the testee data record on the server and made available to the proctor in a processed form for control purposes.

Instructions, previously read by the proctor, are now presented by delta modulation, stored locally, and administered by means of the flow process control and the headset.

Besides the AT-compatible system unit with an arithmetic coprocessor, a 40MB hard disk, Ethernet card, and an audio card for the voice output, the testee station consists of a 14" graphic color monitor with a resolution of 1024 x 768 pixels, a headset, and a specially designed keyboard. The testee keyboard is a high-quality membrane keyboard with 3 x 6 short-travel keys, arranged in an array. Above the top key row and on the side are 7 x 9 dot matrix displays that show the current assignment of the key columns or rows. After switch-over by pressing a key the alternate alphanumerical keyboard assignment is selected. This assignment is shown by illuminated letters above the keys in question. The test-specifically active keys are illuminated, and, an LED illuminates in the actuated key. This virtual keyboard concept enables the fulfillment of the input requirements for all current and future test methods. For motoric tests a joy stick is additionally on the keyboard.

The testee station software, developed in PASCAL like the entire application software, incorporates a module for adaptive testing which estimates local person parameters via an unconditional maximum likelihood approach starting out from given item difficulty estimates based on the Rasch model. On the basis of local estimates, those items from the item pool are extracted and presented which produce maximum gain of statistical test information. The test is continued until it is terminated by a stopping rule (see Habon, MTA 1987).

Particular emphasis was laid on the development of a test design tool box. This unit enables the setup and modification of items, subtests, batteries, audio files, time limits, breakoff criteria, the allocation of parameters to the adaptive algorithm etc. by the user.

This complex tool was developed to take account of the philosophy that the CAT system shall be a flexible, living system which can be adapted by the user to changing requirements during the life cycle of the system.

After a development time of one year only, the exemplary cooperation between engineers and the CAT team of the Psychological Service will lead to the fundamentals of a flexible and effective supporting system, accepted by the users.

Later on, the main system shall be complemented by psychodiagnostic expert systems which will offer various and efficient supporting functions in all phases of the diagnostic process from task analysis via test layout and test administration to judgment finding.

Computerized Adaptive Testing: The CAT-ASVAB Program

by
W. A. Sands *
Officer-in-Charge
CAT-ASVAB Program

Testing Systems Department
Navy Personnel Research and Development Center
San Diego, California 92152-6800

INTRODUCTION

In 1979, a Joint-Service program was initiated for developing a Computerized Adaptive Testing (CAT) version of the Armed Services Vocational Aptitude Battery (ASVAB) for the U.S. Department of Defense.

CAT-ASVAB Program Objectives

The CAT-ASVAB Program has three objectives: (1) to develop a Computerized Adaptive Testing version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB), (2) to develop a microcomputer-based delivery system, and (3) to evaluate CAT-ASVAB as a potential replacement for the paper-and-pencil version of the battery (P&P-ASVAB).

ASVAB Description

The Armed Services Vocational Aptitude Battery (ASVAB) is used by all the Services for initial qualification and classification decisions for enlisted applicants. The battery consists of ten tests: eight power tests and two speeded tests. There are six parallel forms to reduce the potential for compromise. Administration of the battery takes approximately three hours. The paper-and-pencil version (P&P-ASVAB) is administered over one million times each year.

The tests in the current version of ASVAB are: (1) General Science, (2) Arithmetic Reasoning, (3) Word Knowledge, (4) Paragraph Comprehension, (5) Numerical Operations, (6) Coding Speed, (7) Auto & Shop Information, (8) Mathematics Knowledge, (9) Mechanical Comprehension, and (10) Electronics Information.

Conventional vs. Adaptive Testing

Obviously, as the name suggests, Computerized Adaptive Testing (CAT) differs from paper-and-pencil testing in the mode of administration. Less obvious is the way in which items are selected for administration. In the typical, conventionally-administered, paper-and-pencil test, all examinees are administered the same items, frequently in the same sequence.

* The opinions expressed in this paper are those of the author, are not official, and do not necessarily reflect the views of the Navy Department.

The procedure for selecting items in a CAT administration is much different. At the beginning of the test, there is no information about an examinee's ability, so it is assumed to be average. Hence, an item of medium difficulty is chosen for administration. If the examinee answers correctly, the ability estimate is updated (in this case to a higher level). A second item, appropriate to this new estimated ability level, is selected for administration. If the examinee answers this second item incorrectly, the ability estimate is again updated (this time to a lower level). A third item appropriate for this new estimated ability level is selected for administration. The examinee responds with the correct answer, and the ability estimate is again updated. This procedure of selecting an item, administering the item, scoring the response, and updating the ability estimate continues until some stopping rule is satisfied. In fixed-length testing, the stopping rule is satisfied when a specified number of items have been administered. In variable-length testing, items are administered until the ability estimate reaches some target level of measurement precision. Obviously, a hybrid stopping rule can be used. In summary, the test is dynamically tailored to the individual examinee, during the course of the test administration.

The adaptive nature of CAT makes for a very efficient utilization of test items. In a conventional test, all examinees receive all test items, regardless of ability level. This is very inefficient. Low-ability examinees receive many difficult items and simply guess at the correct response. High-ability examinees receive many items that are too easy. Aside from the obvious waste of administration time, this procedure potentially introduces boredom and careless responses.

In the adaptive test, high-ability examinees receive only items that are relatively difficult. Average-ability examinees receive items of average difficulty, while low-ability examinees receive only relatively easy items. This focusing procedure produces a significant reduction in test administration time.

ACAP FIELD ACTIVITIES

The Accelerated CAT-ASVAB Project (ACAP) involves six field activities: (1) Pre-Test, (2) Medium of Administration, (3) Cross-Correlation, (4) Preliminary Operational Check, (5) Score Equating Development, and (6) Score Equating Verification.

Pre-Test

The Pre-Test was designed to evaluate the human-computer system interaction. The ACAP battery was administered to 231 military recruits and 73 high school students, representing the full range of mental ability. Some students were obtained from high school special education classes to bolster the sample of low-ability examinees.

Each of the participants took a questionnaire upon completion of the ACAP battery. This instrument was designed to obtain information on instruction comprehension, fatigue, etc. Between four and eight examinees from each testing session were selected for an in-depth interview. The selection was done to ensure adequate representation of persons finishing both rapidly and slowly.

Results were encouraging. The examinees found the CAT ASVAB faster and easier than paper-and-pencil tests they had taken. They liked the fact that it was self-paced, and that less writing was involved than in the ordinary, paper-and-pencil test.

Some persons expressed a dislike for the fact that they could not go back and change their answers after moving to another item. A few examinees indicated that their eyes became tired. The Pre-Test was completed in November 1986. Based upon information from the questionnaire and interviews, the administration instructions were revised, reducing the reading grade level from the eighth to the sixth grade level.

Medium of Administration

The Medium of Administration study is designed to evaluate the effect of the calibration medium of administration on score precision. The subjects were recruits at the Navy Recruit Training Center in San Diego. Forty-item conventional tests were constructed for General Science, Arithmetic Reasoning, Word Knowledge, and Shop Information. Persons were randomly assigned to one of three groups. The first group was administered the tests on computer. Their data were used to obtain a computer-based calibration of items. The second group took the test in a paper-and-pencil mode. Their results were used to obtain paper-and-pencil calibration information. Each of these calibrations was used to estimate the ability of examinees assigned to the third group of examinees, who took the items on computer.

Data collection for the first phase (involving the four tests) was completed. The number of examinees taking the battery on the computer was 1989, while 983 examinees took the paper-and-pencil version. Analyses are currently underway. The same procedures will be followed for Paragraph Comprehension or Mechanical Comprehension in the second phase.

Cross-Correlation

The Cross-Correlation study is designed to compare the precision of CAT-ASVAB and P&P-ASVAB. The goal was to test 1250 recruits from the Navy Recruit Training Center in San Diego. The operational P&P-ASVAB was one of the following forms: 11A, 11B, 12A, 12B, 13A, or 13B. There were two forms of CAT-ASVAB (non-operational). Finally, there were two non-operational P&P-ASVAB forms employed: 9B and 10B. The operational P&P-ASVAB was the battery that the examinees had taken to enlist in the Navy. The first group took CAT-ASVAB Form 1, then CAT-ASVAB Form 2. The second group took P&P-ASVAB Form 9B, then P&P-ASVAB Form 10B. In each case, the second test was administered about five weeks after the first non-operational test.

The first test phase has been completed, with 1093 examinees taking the two non-operational CAT-ASVAB forms and 1070 examinees taking the non-operational P&P-ASVAB forms. Subsequently, in the retest phase, 786 persons took the CAT-ASVAB, while 761 took the P&P-ASVAB. The database for this study is presently under construction and analyses are scheduled to begin shortly.

Preliminary Operational Check

The Preliminary Operational Check was designed to demonstrate the communications interface between the Accelerated CAT-ASVAB Project (ACAP) System and the U.S. Military Entrance Processing Command (USMEPCOM) System. The test took place at the Seattle Military Entrance Processing Station (MEPS).

The testing procedures were performed jointly by personnel from the Navy Personnel Research and Development Center (NPRDC) and USMEPCOM. Data from 31 examinees, tested in five different sessions were used in the study. These data were

loaded onto the Data Handling Computer at the MEPS, then transferred to the MEPS System-80 minicomputer. Comparison of the data before and after transfer showed the test was completed with perfect accuracy.

Future plans involve merging and editing ACAP results on a MEPS System-80, then telecommunicating the information to USMEPCOM Headquarters.

Score Equating Development

The Score Equating Development study is designed to equate CAT-ASVAB with P&P-ASVAB. Subjects were applicants for enlistment at six MEPS and their satellite Mobile Examining Team Sites (METS). The following six MEPS/METS complexes were selected to be representative of the nation as a whole: San Diego, Richmond, Seattle, Boston, Omaha, and Jackson.

The operational measures included P&P-ASVAB Forms 10A, 10B, 11A, 11B, 13A, and 13B. There were two forms of the CAT-ASVAB (both non-operational). Finally, P&P-ASVAB Form 8A was used as the non-operational reference battery. Subjects were assigned to one of three groups. The first group took CAT-ASVAB Form 1, then the operational P&P-ASVAB. The second group took CAT-ASVAB Form 2, then the operational P&P-ASVAB. The last group took the reference battery (P&P-ASVAB 8A), then the operational P&P-ASVAB. In each case, the testing was done on the same day, or on successive days.

As of 20 November 1988, testing was completed in San Diego (N=618), Richmond (N=1965), Seattle (N=1270), Boston (N=1868), Omaha (N=914), and Jackson (N=1021). Additional data collection was initiated in San Diego on 1 November 1988 to augment the sample collected earlier (N=368).

Results to date have been encouraging. The microcomputer delivery system has performed satisfactorily, exhibiting fewer problems than anticipated. The logistics of administering the battery in the numerous, heterogeneous testing sites has presented no problems which have not been overcome.

Score Equating Verification

The Score Equating Verification study is designed to evaluate the effect of examinee motivation upon item calibration and equating. The subjects will be applicants for military service coming through the same six MEPS/METS complexes used in the Score Equating Development study. The measures will include two forms of CAT-ASVAB and one form of P&P-ASVAB (8A). The CAT-ASVAB scores will be based upon the Score Equating Development study. An equipercentile equating will be performed for subsequent operational use.

The planned schedule for score equating verification involves starting San Diego data collection in February 1990. Scheduled completion date for data collection in Jackson is April 1991.

FUTURE TESTS

The Air Force, Army, and Navy have been supporting R&D for new computerized tests designed to measure abilities untapped by the tests in the current

version of ASVAB. In their Learning Abilities Measurement Program (LAMP), the Air Force has investigated a number of memory, reasoning, learning, and reaction time tests. The Army's Project A covers a broad list of abilities and includes psychomotor tests requiring special input devices (e.g., joysticks). Navy R&D has focused primarily on computerized tests of cognitive speed and spatial ability.

By measuring domains not measured by the current ASVAB tests, these new tests offer the promise of incremental validity. Even a small increase in validity for personnel selection would translate to a substantial financial savings for the military services because of the large number of persons tested each year. These savings could be used to offset the cost of procuring, deploying, and operating a nationwide microcomputer-based personnel accessioning system.

Future Tests - Design for Validation in Ten Navy Schools

John H. Wolfe

Navy Personnel Research and Development Center

Introduction

The Armed Services have been engaged in a joint services project to develop an operational demonstration system for administering a computerized adaptive version of the Armed Services Vocational Aptitude Battery (CAT-ASVAB). At the same time, new experimental computerized tests have been developed that measure abilities not covered by the ASVAB, e.g. tests of working memory, spatial visualization, psychomotor abilities, and reaction time.

In order to show that such "future tests" are useful, it is not only necessary to validate them in the usual ways, but also to show that they have incremental validity when added to the ASVAB or CAT-ASVAB.

The measurement of incremental validity requires investment in careful research design, experimental testing facilities, and large sample sizes. If the incremental validity of future tests can be demonstrated, their use could save the Armed Services hundreds of millions of dollars annually in manpower costs and improved performance.

The Test Battery

After collecting data on reliabilities and correlations with ASVAB for over thirty experimental tests, we have decided to focus on tests of spatial ability, mental speed, and working memory. The tests chosen for validation include the following:

- a. The Integrating Details test - a measure of spatial visualization ability (Alderton, in press).
- b. Assembling Objects - a spatial visualization test developed by the Army Project A.
- c. Spatial Relations from the ASVAB Form 6 administered by computer
- d. Mental Counters - a test of working memory and cognitive speed (Larson, Merritt, & Williams, 1988).
- e. Spatial Reasoning - a nonverbal reasoning test developed by the Army Project A.
- f. Perceptual Speed - detecting same or different letters or figures.
- g. The Armed Services Applicant Profile (ASAP) - a biographical data questionnaire (Trent, in press).
- h. A test-taking motivation and fatigue questionnaire.

The primary hypothesis to be tested is that the accuracy scores for tests *a* to *f* will increase the validity of the ASVAB when added to it in a multiple regression. Tests *a*, *c*, *d*, and *f* also generate one or more latency scores. These will be tested for incremental validity in a later, exploratory analysis.

The opinions expressed in this paper are those of the author, are not official, and do not necessarily reflect the views of the Navy Department.

The ASAP, which was developed to predict attrition rather than performance, will be validated against non-academic attrition. It will also be used in a later exploratory analysis for predicting academic performance.

The test-taking motivation and fatigue questionnaire will be used to screen out subjects who are not doing their best on the experimental tests. The latency measures will be used for the same purpose, i.e., examinees who press keys as fast as they can with low accuracy will be rejected from the sample.

A controversial issue is whether it is necessary to readminister the ASVAB at the same time the experimental tests are given in order to establish incremental validity. Data from the civilian world have generally shown increased validity for aptitude tests the closer that they are given to the criterion in time (Henry & Hulin, 1987). There is some unpublished information, however, that indicates that validities of the ASVAB given at Recruit Training Centers are no better than pre-enlistment ASVAB validities for predicting "A" school training performance, and that retests are only more valid when given concurrently with the criterion. Since this particular study is predictive rather than concurrent, we have chosen to readminister the ASVAB or CAT-ASVAB only if additional funding is provided to do so.

Criteria

Traditionally, the ASVAB has been validated against school performance in enlisted technical training. We have some studies under way to validate new tests against job performance criteria, but in the present study, we will continue to use "A" school final grades or time to completion as the criterion.

Schools Selected

In selecting the schools for validation, we looked for schools with high flow rates, high attrition, and where spatial ability might be related to performance. In the end, it was necessary to concentrate on schools with large flow rates so that the data collection could be completed as quickly as possible. The candidate list of schools is

AC	Air Traffic Controller
AD	Aviation Machinist's Mate
AE	Aviation Electrician's Mate
AMS	Aviation Structural Mechanic
AO	Aviation Ordnanceman
AT/AQ/AX	Aviation Electronics Technician
FC	Fire Control Technician
MM	Machinist's Mate
OS	Operations Specialist
HM	Hospital Corpsman

Subjects

The examinees will be recruits in their first or second week of basic training at Navy Recruit Training Centers who have school training guarantees for the ratings of interest. There are several advantages to using recruits over going to the "A" schools directly:

- a. The validities will be predictive validities, not concurrent validities. Some people have argued that spatial ability, for example, is acquired through experience in dealing with spatial problem solving tasks, such as those involved in repair of mechanical equipment. If this theory is true, then concurrent validities of spatial aptitude tests would be misleading estimates of predictive validities.

- b. Pragmatic factors involved in scheduling subjects make it much easier to test recruits at Recruit Training Centers than at Navy "A" schools.

Statistical Design

The basic statistical test to be used is the F-test for comparing full and restricted models in regression analysis.

Suppose that a criterion is predicted by a multiple regression equation from m conventional tests, and suppose that we wish to test the significance of adding k new predictors to the battery. The appropriate significance test is

$$F_{k, n-m-k-1} = \frac{(R^2_{m+k} - R^2_m)/k}{(1 - R^2_{m+k})/(n-m-k-1)}, \quad (1)$$

where R_m = the multiple correlation of the m predictors with the criterion, R_{m+k} = the multiple correlation of the $m+k$ predictors with the criterion, and n = the sample size.

The quantity

$$f^2 = \frac{(R^2_{m+k} - R^2_m)}{(1 - R^2_{m+k})} \quad (2)$$

is often used as a measure of *effect size*.

Number of ASVAB Variables

There are a number of controversial issues in applying this statistical test. The first issue is how many variables should represent the ASVAB in the regression equation, i.e., what should the value of m be. The ASVAB could be represented by all ten subtests, by factor scores on four major factors, or by g , a general ability factor score.

It has been argued that regression with all ten subtests of the ASVAB would lead to sampling errors because of the high collinearity of the subtests. I believe this argument is incorrect, because collinearity affects the regression weights only. The estimates of multiple R are not affected by collinearity of the predictors, and the F-tests are also unaffected. This is easy to see, because with principal components analysis, any predictor set could be transformed into an equivalent one that has uncorrelated predictors but the same multiple R and the same predicted values of the dependent variable.

Schmidt, Hunter, and Larson (1988) examined the effectiveness of three types of ASVAB prediction equations using ten subtests, three factor scores, or a single g for predicting school performance in ten Navy ratings. Sample sizes ranged from 928 to 2598. They found that, on the average, the multiple R from using three factors was 2% higher than from using g , and the multiple R from using all ten subtests was 8% higher than from using g alone.

These results have an important implication for the design of an incremental validity study. If we were to use a g representation for the ASVAB, then a "new" test that was merely a parallel form of an ASVAB subtest might appear to have incremental validity. There is no alternative, then, to using all ten subtests of the ASVAB as the basis for our restricted model.

Number of New Predictors

Notice that Equation (1) uses all of the new predictors at once. It is common practice to evaluate each predictor one at a time, so that $k = 1$ in the F-test, making it equivalent to a t -test. The trouble with multiple t -tests, is, of course, that if you make lots of them, sooner or later one of them is going to appear significant, just by chance. Bonferroni and others have proposed fixes

to this situation by setting up higher levels of significance for each t -test individually so that the overall significance level is the desired value of, say, .05 or .01, but such methods tend to lack power (Alt, 1982).

Another approach might be stepwise regression analysis, where k gradually increases. As in the case of multiple t tests, the significance tests printed out by all commonly used statistical packages are wrong. Wilkinson (1979) presents tables for estimating the significance of the overall multiple R in stepwise regression, but I know of no good ways of estimating the significance of the increments in stepwise regression.

Cohen and Cohen (1983, pp. 172-176) recommend a procedure analogous to Fisher's protected t -test in analysis of variance. First, the significance of the multiple R with all variables in the regression equation is tested. If this proves significant, then one can safely make t -tests one at a time for the individual variables in the equation. This is the approach adopted for this design.

Power Analysis for Sample Size

Schmidt, Hunter, and Dunn (1987) estimate that adding new perceptual tests to the ASVAB would increase its validity from .59 to .61. By Eq. (2), the effect size is .0382. Using Eq. 4.5.2 from Cohen and Cohen (1983), the following table can be constructed:

Number of Cases Needed for 90% Chance of Detecting Incremental Validity		
k	Significance Level, α	
	.05	.01
6	473	624
10	558	727

This tells us that in order to detect an incremental validity with the effect size of .0382 with a 90% chance of success, using a significance level of .05, we should test 473 examinees if only the accuracy scores are used, and 558 examinees if four additional latency scores are used.

On the average, we can expect about 25% attrition from the first week of recruit training to the completion of "A" school. Also, past experience seems to indicate that another 25% of the examinees will engage in random key-pressing behavior under unmotivated conditions. Thus the figures in the above table should be increased by about 78% in order to obtain adequate sample sizes for validation. In this particular validation study, we plan to test about 1000 subjects in each rating.

Corrected Correlations

It is possible that a new predictor might have incremental validity for the wrong reasons. For example, it might have so much overlap with the ASVAB that its inclusion in the predictor set merely increases the reliability of the ASVAB composite, and therefore its validity. An equivalent result could be obtained by retesting with a parallel form of the ASVAB and combining the results of the two testings. In order to rule out this possibility, the correlations of the ASVAB subtests with the criterion and with the new predictors should be corrected for attenuation of the ASVAB subtests. It is unnecessary to correct for attenuation in the criterion or the new predictors in order to show that the incremental validity is not due to ASVAB unreliability. Also, for the purpose of testing the significance of incremental validity, it is unnecessary and undesirable to correct for restriction in range. However, before correcting for attenuation, it may be necessary to correct the reliability coefficients so that they are appropriate for the restricted

sample.

Selector Composites with New Predictors

Unfortunately, the Navy uses integer-weighted composites, not regression equations for selection and classification into technical training. This means that all of the above analyses, including significance tests for incremental validity, will have to be redone for composites. The regression-based analysis is still necessary, because we do not want to add new predictors when merely re-weighting the ASVAB would accomplish the same improvement and make the new predictor unnecessary. Nevertheless, a new predictor that has incremental validity for least-squares regression may not improve any existing selector composite.

There appear to be a large number of ways to proceed with constructing selector composites. Possibly the simplest is to use stepwise regression with all ASVAB and new tests in the pool of potential predictors in order to identify the tests in the composite, then determine the validity for the unit-weighted composite. Finally, use Eqn. (1) to estimate significance of the difference between the validities of the new composite and the old composite. If the old composite is a subset of the new one, then $m = 1$, since the old composite is a single prespecified variable; otherwise $m = 10$.

Utility Analysis

Once a set of predictors has been found to have incremental validity in regression analysis and in selector composites, one can proceed to do a cost/benefit analysis of adding the new tests to the ASVAB. At this point, shrinking the multiple Rs using Wherry's formula, correcting the criterion (but not the predictors) for unreliability, and doing a multivariate correction for restriction in range will be necessary to obtain an estimate of the magnitude of the improvement. Schmidt, Hunter, and Dunn (1987) have shown that a 3% improvement in validity (from .59 to .61) can eventually save the Navy \$83 million annually in improved performance. Based on some small sample studies, we expect the improvement to be much larger than this, on the order of 10% for some ratings. Therefore, the potential payoff from this research could be very large.

Summary

To prove that the addition of new tests to the ASVAB can increase validity is a challenging undertaking. Because we are looking for small improvements, on the order of 3% or more, large samples are required. The statistical approach must be sound and insure that artifacts do not creep into the results. This paper has addressed potential problems arising from not retesting with the ASVAB, outliers and unmotivated examinees, ASVAB unreliability, collinearity in regression, the number of variables in regression, multiple *t*-tests, statistical power, corrections for attenuation and restriction in range, and use of selector composites. We welcome any criticism or suggestions for improving this design.

REFERENCES

- Alderton, D. L. (In press). *Development and evaluation of Integrating Details: A complex spatial problem solving test*. San Diego: Navy Personnel Research and Development Center.
- Alt, F. B. (1982). Bonferroni inequalities and intervals. In S. Kotz, N. L. Johnson, & C. B. Read (Eds.) *Encyclopedia of statistical sciences*, 1, 294-300.
- Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

- Larson, G. E., Merritt, C. R., & Williams, S. E. (1988). Information processing and intelligence: some implications of task complexity. *Intelligence*, 12, 131-147.
- Hem, J. R. A., & Hulin, C. L. (1987). Stability of skilled performance across time: some generalizations and limitations on utilities. *Journal of Applied Psychology*, 72, 457-462.
- Schmidt, F. L., Hunter, J. E., & Dunn, W. L. (1987). *Potential utility increases from adding new tests to the Armed Services Vocational Aptitude Battery (ASVAB)*. Unpublished manuscript (Battelle Contract Delivery Order 53) submitted to Navy Personnel Research and Development Center.
- Schmidt, F. L., Hunter, J. E., & Larson, M. (1988). *General cognitive ability vs. general and specific aptitudes in the prediction of training performance: some preliminary findings*. Unpublished manuscript (Battelle Contract Delivery Order 53) submitted to Navy Personnel Research and Development Center.
- Trent, T. (in press). *Joint services adaptability screening: validation of the Armed Services Applicant Profile (ASAP)*. NPRDC TR 88-8. San Diego: Navy Personnel Research and Development Center.
- Wilkinson, L. (1979). Tests of significance in stepwise regression. *Psychological Bulletin*, 86, 168-174.

VALIDATION OF THE AIR FORCE RESERVE OFFICER TRAINING CORPS SELECTION SYSTEM

Linda R. Elliott
Air Force Human Resources Laboratory
Brooks Air Force Base, TX 78235-5601

This study examines the assessment process used to select candidates for Air Force Reserve Officer Training Corps (AFROTC) Professional Officer Course (POC). The current AFROTC selection system assigns a Quality Index Score (QIS) to each POC candidate, based on scores from six predictor variables, which are weighted and combined to form an overall score, or QIS (Jackson & Gordon, 1977). QIS scores were examined for their relation to several training and job performance measures. The predictive validity of QIS scores as presently calculated (6 factors) was compared to that of QIS scores calculated from a previously used 10-factor formula and a proposed 3-factor formula. In addition, QIS formula weights were compared to the optimal weights assigned to each variable in regression analysis.

METHOD

Subjects and Criteria. Data on AFROTC applicants from fiscal years 1978-81 were obtained from Headquarters AFROTC (N = 13,722). The majority of subjects were caucasian (82%) males (81%). Samples were generated on the basis of nine criteria: a) POC selection (POC SEL); b) POC completion (POC COM); c) POC instructor ratings (POC RAT; 1-5 scale); d) POC distinguished graduates (POC DG); e) technical school final grades (TS GRD); f) experimental supervisory ratings of job performance (EXP JOB, 1-9 scale); g) experimental ratings of potential for career progression (EXP PCT, 1-9 scale); h) experimental ratings of motivation (EXP MOT, 1-9 scale); and i) Officer Effectiveness Reports (OERS, 1-6 scale, 1 = superior performance).

Predictor Variables. Eleven predictor variables have been used by AFROTC for the determination of QIS scores: a) Detachment Commander's rating (DCR) (an overall rating of acceptability of an applicant, ranging from 0-8); b) Cadet Rank (CR) (assigned rank from 1 (most desirable) to 50 (last choice)); c) Total Cadets Ranked (TCR) (the actual number of applicants ranked); d) General Military Course Credit (GMC) (indicates whether the applicant had prior military coursework or service); e) AFROTC grade point average (AFROTC GPA) (prior military experience counted as 3.0 AFROTC GPA); f) Technical Credit (TC) (credit for technical coursework completion); g) Air Force Officer Qualifying Test (AFOQT): Academic Aptitude Composite; h) AFOQT: Quantitative Composite; i) AFOQT: Verbal Composite; j) Scholastic Aptitude Test (SAT) scores; and k) Cumulative grade point average (GPA).

Variables which have been used to compute QIS scores are shown in Table 1, together with their operational weights.

Table 1. Weights Used in Computation of QIS^a

Variables	Range	Variable Weight		
		1978 - 1982 QIS	Current QIS	Proposed QIS
1. Det. Commander Rating	(0 - 8)	1.9625	3.8233	2.0000
2. Cadet Rank	(1 - 50)	-.1106	-	-
3. Total Cadets Ranked	(1 - 131)	.0362	-	-
4. Gen. Military Crse Credit	(0 - 1)	1.5125	-	-
5. AFROTC GPA	(0 - 400)	.0157	-	-
6. Technical Credit	(0 - 1)	2.1332	-	-
7. AFOQT-Academic Aptitude	(1 - 99)	.1687	.1293	2.0000
8. AFOQT-Quantitative	(1 - 99)	.0556	.1125	-
9. AFOQT-Verbal	(1 - 99)	-	.1189	-
10. SAT	(0 - 1600)	.0225	.0187	-
11. Cumulative GPA	(0 - 400)	.0931	.0719	.6000

^aWeights are applied to actual predictor scores; their relative magnitude cannot be directly compared without consideration of the predictor metric.

Analyses. Three QIS scores were calculated for each AFROTC applicant using the 3, 6, and 10 variable formulae. Descriptive statistics for the individual predictor variables, the three types of QIS scores and nine criteria were computed and are available upon request. This descriptive data will also be available in a technical report (Cowan, Elliott, & Wegner, in review). The correlations between each type of QIS score and the training and job performance criteria were examined for significant differences in predictive validity.

The same variables which comprise the three types of QIS scores were examined using regression analyses to determine if different variable weights would result in significantly higher correlations with criteria. This resulted in a separate set of optimal weights for the prediction of each criterion. The correlations between QIS scores and each criteria were compared to the multiple correlations obtained from each corresponding regression model. Differences between the correlations were tested for significance.

III. RESULTS

Relation between individual predictor variables and criteria. Correlations between the 11 individual predictor variables and each criterion are shown in Table 2. Also listed in Table 2 are the correlations for each type of QIS score (variables 12-14) and the multiple correlations obtained for each regression model (variables 15-18). Correlations were not corrected for reliability and restriction in range of the predictor variables, so that reported correlations are conservative estimates. Also, the variance of the job performance criteria was very restricted, therefore the low correlations are not surprising.

Table 2. Correlations Between QIS Variables and Criteria

Predictor variables	Criteria								
	POC SEL	POC RAT	POC COM	POC DG	TS GRD	EXP JOB	EXP POT	EXP MOT	OER
Individual variables (zero-order correlations)									
1. D.C. Rating	.23**	.19**	.11**	.21**	-.00	.10**	.11**	.11**	-.06*
2. Cadet Rank	-.19**	-.19**	-.11**	-.22**	-.02	-.03	-.02	-.02	.01
3. Total Ranked	.04	-.01	.01	.01	.02	.07*	.10**	.10**	-.05
4. Gen Mil Credit	.09**	.05	.03	.04	.12**	.06*	.06*	.05	-.06*
5. AFROTC GPA	.15**	.18**	.10**	.19**	.09**	.06*	.10**	.07*	-.06*
6. Technical Credit	.03	.04	.01	.06*	.16**	.08**	.11**	.06*	-.12**
7. AFQOT-AA	.10**	.12**	.06*	.15**	.37**	.07*	.14**	.05	.09**
8. AFQOT-Quant	.10**	.08**	.07*	.12**	.33**	.05	.12**	.03	-.11**
9. AFQOT-Verbal	.08**	.13**	.04	.13**	.33**	.03	.06*	.01	-.04
10. SAT	.11**	.12**	.07*	.16**	.39**	.06*	.13**	.04	-.09**
11. Cumulative GPA	.13**	.24**	.17**	.31**	.20**	.07*	.06*	.04	-.05
QIS Scores (zero-order correlations)									
12. 10-QIS-Assigned	.19**	.23**	.15**	.29**	.41**	.11**	.16**	.08**	-.12**
13. 6-QIS-Assigned	.21**	.24**	.14**	.28**	.38**	.10**	.16**	.08**	-.11**
14. 3-QIS-Assigned	.15**	.22**	.14**	.27**	.40**	.10**	.14**	.06*	-.10**
Regression Scores (multiple correlations)									
15. 11 variables	.27**	.30**	.20**	.36**	.45**	.17**	.22**	.18**	-.16**
16. 10-variables	.27**	.30**	.20**	.36**	.44**	.16**	.21**	.17**	-.16**
17. 6-variables	.25**	.29**	.19**	.35**	.43**	.13**	.19**	.13**	-.12**
18. 3-variables	.25**	.28**	.18**	.35**	.40**	.12**	.17**	.11**	-.11**
N	13,722	5,249	9,450	7,679	1,645	1,082	1,080	1,080	3,923

*significant at .05 level; **significant at .01 level.

Comparison of the predictive validity of each type of QIS score. Table 2 displays the zero-order correlation coefficients between the three types of QIS scores (variables 12-14) and the nine criteria. Correlations were highest for the technical school grades ($r = .38$ to $.41$) and were lowest for the job performance criteria ($r = .06$ to $.16$). Note that correlations are negative for the OER criterion, because the OER uses a reverse rating scale, from 1 (highest rating) to 6 (lowest rating).

Differences between correlation coefficients of each type of QIS score were tested for significance using Hotelling's formula (Guilford & Fruchter, page 164, 1978). While there are a few differences between the QIS methods for predicting training and job performance which were statistically significant, the differences are not large. The largest difference for predicting post-selection criteria was between the 10 and 6 factor QIS correlations with technical school grades, which differed by .03. Otherwise, all other differences in correlations are less than .02.

Predictive validity of regression models. Table 2 provides a summary comparing the multiple correlation of each of the three regression models (variables 15-18) for the prediction of each nine criteria. F tests were calculated to determine if differences between the regression models were statistically significant. There were significant differences between model 1 (10 variables) and model 2 (6 variables) for all criteria except the technical school grades. However, the differences were relatively small. There was very little difference in the predictive validity of model 2 (6 variables) compared to model 3 (3 variables); differences were nonsignificant for all criteria but three: POC instructor ratings, technical school grades, and OER ratings. The difference in multiple correlation values between model 2 and model 3 for these three criteria were also relatively small.

Results indicate that the deletion of three variables from the current method of QIS determination will not practically impair the predictive validity of the QIS scores. This is as expected, since the three variables which have been suggested to be deleted, the SAT, AFQOT-Verbal, and AFQOT-Quantitative scores, are highly correlated with one of the remaining variables - the AFQOT Academic Aptitude score ($r = .85, .83, \text{ and } .82$, respectively). Intercorrelations between the 11 QIS predictor variables are available upon request.

Predictive validity of QIS weights versus their corresponding regression weights. The multiple correlations obtained from regression analyses of model 1, model 2, and model 3 were compared to the correlations obtained for the 3-, 6-, and 10-factor QIS scores for each criterion, using F tests of statistical significance. Correlations were significantly higher for the regression models as compared to the corresponding QIS methods for all comparisons but one (exception = 6-factor QIS versus model 2 for the prediction of experimental measures of job performance). Differences in the predictive ability of the QIS scores and the regression models should be interpreted with some caution, since statistical analysis procedures capitalize on chance. The multiple correlations obtained in these analyses may shrink if the regression weights are applied to a second group of subjects. Regression weights are expected to remain significantly more predictive of criteria that had a larger difference in prediction between QIS scores and corresponding regression models, such as performance in POC training, POC completion, POC distinguished graduates, and experimental measures of motivation and potential for career progression.

Relative magnitude of QIS weights vs regression weights. Regression weights for each variable differed according to the criterion being predicted. In order to compare the QIS weights to the regression weights, single regression weights were computed for each variable by averaging across the eight training and performance criteria. Table 3 lists the averaged regression weights for each of the three models, along with the assigned weights for each QIS formula. The weights were rescaled with the weight for AFQOT Academic Aptitude set to 1, to facilitate direct comparison of weights.

When the 10 factor QIS method is compared to regression model 1 it is noted that model 1 increases the weight ratio markedly for the Detachment Commander's Rating, the General Military Course Credit, and Technical Credit. Comparing the current 6-factor QIS to regression model 2 reveals that the regression weight for Detachment Commander's rating is twice the assigned QIS weight. Regression weights for AFQOT-Quantitative score and Cumulative Grade Point Average are also higher than the operational QIS weights. For the proposed 3-factor QIS and model 3 comparison, it was found that the

Table 3. Relative Magnitude of QIS weights versus Raw Score Regression Weights^a

Variables	Weights					
	1978-82		Current		Proposed	
	QIS	Mod 1	QIS	Mod 2	QIS	Mod 3
1. Det Commander Rating	11.63	17.02	29.57	64.31	1.00	6.91
2. Cadet Rank	-.66	-.07	-	-	-	-
3. Total Cadets Ranked	.27	.46	-	-	-	-
4. Gen. Mfl. Credit	8.97	62.83	-	-	-	-
5. AFROTC GPA	.09	.13	-	-	-	-
6. Technical Credit	12.64	37.81	-	-	-	-
7. AFQOT-AA	1.00	1.00	1.00	1.00	1.00	1.00
8. AFQOT-Quantitative	.33	.31	.87	3.00	-	-
9. AFQOT-Verbal	-	-	.92	1.63	-	-
10. SAT score	.13	.24	.14	.88	-	-
11. Cumulative GPA	.55	.65	.56	2.25	.30	.25

^a The weight for AFQOT-Academic Aptitude was set to 1 across all models, so that assigned and regression weights can be directly compared.

regression model assigned a lower weight for the Cumulative Grade Point Average and a higher weight to the Detachment Commander's rating. It is noted that each of the regression models shows a marked increase in weight for the Detachment Commander's Rating over the three QIS ratios.

The averaged raw regression weights resulted in higher weights for Detachment Commander ratings in all models, when compared to the QIS weights. In the 6-factor model, the weight for Detachment Commander ratings was doubled, and in the 3 factor model, the weight was increased sixfold. It should be noted that these raw regression weights do not represent the actual ratio of differences in importance of the predictor variables, because they do not control for differences between variables in their metric (i.e. mean, range, variability).

IV. DISCUSSION

The current AFROTC selection system was demonstrated to have a significant degree of predictive validity for a variety of training and performance criteria. The 6 factor QIS scores now used for entry into the AFROTC POC were significantly related to all measures of training success and job performance. Predictive validity of this method was highest for technical school final grades and lowest for experimental ratings of job performance, potential for career progression, and motivation.

Scores generated by regression weights were more highly related to selection, training, and job performance criteria than were their corresponding QIS scores. Differences in predictive validity between the QIS scores and their corresponding regression models were greatest for selection and POC criteria and lowest for technical school grades and job performance measures. It should be noted that the predictive validity of the regression weights may shrink somewhat when applied to another group of subjects. However, since differences in predictive validity were substantial for the selection and POC criteria, it is likely that the predictive validity of the regression weights would remain significantly higher when applied to another group of subjects.

The averaged raw regression weights resulted in higher weights for Detachment Commander ratings in all models, when compared to the QIS weights. In the 6-factor model, the weight for Detachment Commander ratings was doubled, and in the 3 factor model, the weight was increased almost sevenfold.

This report was restricted to the validation of variables which are or have been used for AFROTC selection. These variables may or may not be the optimal predictors of training and/or performance criteria. Further research would be necessary to determine the relation of other applicant characteristics to training and job performance criteria, such as previous work experience, prior military service, awards and/or

achievements, and participation in extracurricular activities, such as team sports. These characteristics are probably considered to some extent by the Detachment Commander in the assigned overall ratings. Since these ratings were positively related to performance criteria, there is reason to expect that some of the additional characteristics would be predictive of AF Officer performance. A list of potential predictors could be obtained by eliciting information from Detachment Commanders as to the factors they consider most relevant in their rating decisions. Identification of applicant characteristics which are predictive of AF Officer performance would provide further information to the Detachment Commanders and serve to enhance the reliability and predictive validity of Detachment Commander ratings. At this time, the reliability of the Detachment Commander ratings is not known.

Recently (July 88), a plan for restructuring the AFROTC selection system was submitted in response to a general directive from the Chief of Staff of the Air Force to give greater consideration to candidate attributes such as motivation, officership, and adaptability to military life. The new plan deletes the SAT and AFQOT Academic Aptitude scores and adds several new measures, such as a physical fitness test, a written communication assessment, and a face-to-face board interview. Assessment variables and proposed weights are listed in Table 4.

Table 4. Variables and weights included in recommended ROTC cadet evaluation form for four week evaluation

Assessment variables	Range	Weight	Maximum Score
I Motivation			
Review Board	0 - 24	0.5208	12.4792
Cadet Traits	0 - 27	0.2315	6.2505
Military Aptitude	0 - 15	0.4167	6.2505
Physical Fitness	0 - 400	0.0156	6.2400
Det Commander Rating	0 - 8	3.1250	25.0000
II Ability to Learn			
AFQOT-Verbal	1 - 99	0.2104	20.8296
AFQOT-Quantitative	1 - 99	0.2104	20.8296
GPA	0 - 4	5.2078	20.8312
Total QIS			112.4706

The board interview rating would be based on 6 factors worth 3 points each: a) self confidence, b) human relations, c) oral communication skill, d) written communication skill, e) Air Force interest, and f) extracurricular activities. Cadet traits to be assessed include flexibility, assertiveness, perseverance, responsibility, self confidence, cooperation, decisiveness, organization, and ethical principles. Military aptitude is assessed through AFROTC GPA, military bearing, military knowledge, membership in AFROTC activities, and AFROTC honors. As can be seen in the table, the variables measuring academic ability provide 50% of the total score. In the current system, measures of academic ability provide 75% of the total score. The net result of the new system will be a greater weight for characteristics previously considered only in the Detachment Commander rating. Results from this study indicate that the greater weight should not reduce the predictive validity of QIS scores, and may in fact increase the predictive validity for some criteria.

References

- Alley, W.E., & Gibson, T. A. (1977). Predicting success in the AFROTC scholarship program (AFHRL-TR-77-11 AD-A041-132). Lackland AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory.
- Cowan, D., Elliott, L., & Wegner, T. (In review). Validation of the Air Force Reserve Officer Training Corps Selection System. Brooks AFB, TX: Air Force Human Resource Laboratory.
- Guilford, J.P. & Fruchter, B. (1978). Fundamental Statistics in Psychology and Education. New York: McGraw Hill.
- Jackson, D. K. & Gordon, M. M. (1977). The development of a weighted selection system for the AFROTC Professional Officer Course. AFROTC Education Journal, winter, 1977.
- Stokes, R. W. (1984). Preserving the Lamberent Flame: Traditional values and the USAF officer accession program (AU-ARI-83-8). Maxwell AFB, AL: Air University Press.

The Relationship Between Aptitude and Officer Performance
Dianne C. Brown
The Human Resources Research Organization (HumRRO)

In contrast to enlisted personnel, a variety of aptitude tests are used in the selection of military officers. These tests not only vary by Service, but by officer commissioning program as well, thus complicating the task of analyzing officer aptitude. The Scholastic Aptitude Test (SAT), widely used for college admissions in the civilian sector, is the closest to a consistent aptitude measure of officers. The SAT is used for selection into the majority of officer commissioning programs (i.e., academies, and Reserve Officer Training Corps programs). In addition, because military officers are primarily college graduates, it can be expected that most officers will have taken college entrance examinations, particularly the SAT, regardless of commissioning source. Thus, the SAT seems well suited, for the present study, to serve as the uniform measure of officer aptitude and for comparing officers of varying aptitude levels. Before analyzing the relationship between aptitude and officer performance, this paper first defines officer aptitude levels using SAT scores, documenting the procedures used to accomplish this.

The first step in defining officer aptitude levels was to create a database which contained the necessary information. The Defense Manpower Data Center (DMDC) officer cohort files do not contain aptitude test score information. Therefore, the DMDC file for officers commissioned from FY 1975 through FY 1985 was loaned to the Educational Testing Service (ETS) where SAT score data were matched to individuals via social security numbers. The match rate for SAT scores was approximately 56 percent. In other words, it was possible to obtain SAT scores for 56 percent of DMDC's original officer cohort file. The 44 percent that did not match is likely attributable to some individuals having taken the American College Test (ACT) rather than the SAT; some individuals having taken no college entrance examination; or social security numbers that did not match. Subsequent to matching, social security numbers were erased from the database by ETS. This "SAT/Officer" database, provided to DMDC by ETS, serves as the basis of analyses of high aptitude officers for HumRRO's high aptitude project.¹

Representativeness of the Matched SAT/Officer Sample

To determine whether the resulting matched sample adequately represented the population of military officers, an analysis of its demographic characteristics was performed. The total FY75-85 officer cohort was compared to the 56 percent SAT matched file in terms of gender, race, education, and Service. These data indicated no significant differences between the two cohorts. Therefore, it can be expected that the matched cohort is representative of the total FY75-85 officer cohort.

Officer Aptitude Levels

To define officers in terms of general levels of aptitude for the high aptitude study, a reference population of SAT takers was used to establish a normative base. SAT Verbal and Math scores, provided by ETS, were combined into a single index (analogous to the single Armed Forces Qualification Test

¹ The High Aptitude Project was contracted to HumRRO through the Office of the Assistant Secretary of Defense (Force Management and Personnel).

[AFQT] measure which consists of verbal and quantitative components) and then converted to the "appropriate" percentile range. The general procedures for converting SAT score ranges into percentile score ranges were as follows.

First, an SAT reference population was established. The appropriate normative base was considered to be the SAT taking population for years corresponding to the dates that most FY 1975-1985 officer accessions would have taken the SAT for college admissions. Assuming that most of the officer commissions are recent college graduates, 1970 through 1981 SAT takers would be the ideal referent. However, prior to 1972 the total number of SAT takers was counted as the number of SAT tests taken. In 1972, ETS began counting SAT takers as the total number of graduating seniors who took the test, eliminating the double counting of individuals. To assure consistent data, the reference population chosen for the current project is the 1972-1981 SAT taking population.

The next step consisted of transforming the defined norming group's SAT standard scores into their percentile equivalents. ETS provided total SAT score ranges corresponding to the percentile bands of interest (i.e., percentile ranges similar to AFQT categorizations or 93-100; 65-92; 50-64; 31-49; 10-30; and 1-9² for each year included in the reference population. ETS estimated these SAT total scores within percentile ranges from SAT verbal and math means and estimates of standard deviations. To produce one single percentile scale based on the reference population, the nine years of data were combined. SAT scores for each year were converted to standard z scores, and the number of examinees scoring at each 10 point interval was estimated.² The percentile ranges were computed for the combined 10 years, resulting in the following reference population for defining the officer aptitude levels, as shown in table 1.

Table 1
SAT Score Ranges and Corresponding
Percentile Ranges and Aptitude Categories

SAT Score Range	Percentile Score Range	Aptitude Category	
1220-1600	93-100	0-I	} High
980-1210	65-92	0-II	
880-970	50-64	0-III A	} Average
740-870	31-49	0-III B	
640-730	10-30	0-IV	} Low
400-630	1-9	0-V	

SAT scores for high aptitude officers are defined as those scoring within the 65th through the 100th percentiles on the SAT; average aptitude scoring within the 31st through the 64th percentiles; and low aptitude

² Subsequent to 1971, SAT scores were rounded to the nearest 10. The overwhelming majority of the SAT-taking population and the officer cohort being used in this study have SAT scores divisible by 10. Each 10 point interval was assumed to be linearly distributed.

within the 1st through the 30th percentiles. In other words, high aptitude officers are defined as those scoring at or above 980 on the SAT, with this range set on the basis of the percentile distribution of the above-specified reference population. Average aptitude officers will comprise the 31st through 64th percentiles. Using the SAT/matched database for the FY 1975 through 1985 officer cohort, comparisons can be made of the performance of high and average aptitude officers.

It should be noted here that the SAT and the AFQT were in no way equated. The percentile ranges that are used to define the various aptitude categories of the AFQT were simply "borrowed" and applied to an SAT reference population. These percentile ranges and categories were used simply because they provide a familiar means of identifying and defining aptitude levels for officer personnel. Because the two tests have not been equated, the aptitude categories cannot be considered equivalent. Through self-selection, SAT takers (i.e., the reference population for this study) tend to be among the upper end of high school graduating class rank. This is markedly different than the nationally referenced AFQT, for which there is a much wider range of ability. Therefore, in viewing the data in this paper, what is termed "low" or even "average" aptitude for officer personnel must be considered low or average in reference to a rather select group.

SAT/Matched Officers by Aptitude

Frequencies and percentages of officers in the various SAT categories by Service are shown in Table 2. As might be expected, proportions of low aptitude officers are minimal. Average aptitude officers comprise approximately 27 percent of the SAT/officer cohort, and about 67 percent of the cohort are high aptitude officers. These proportions are consistent with what one would expect in a selected group. The largest proportion of officers in the present SAT/officer cohort (46.7 percent) in category O-II is consistent with typical qualifying SAT scores. Based on 1986 standards, SAT qualifying scores ranged from 850 to 1120 across the Services and commissioning sources.

Table 2
Number and Percent of SAT/Officers Commissioned
FY1975-85 by SAT Aptitude Level and Service*

Service	High Aptitude				Average Aptitude				Low Aptitude				Total
	O-I		O-II		O-IIIA		O-IIIB		O-IV		O-V		
	N	%	N	%	N	%	N	%	N	%	N	%	
Army	9,817	17.0	24,328	42.2	8,976	15.6	8,486	14.7	3,538	6.1	2,539	4.4	57,684
Navy	12,176	27.3	22,858	51.2	5,371	12.0	3,280	7.3	738	1.7	253	0.6	44,676
Marine Corps	1,711	13.1	6,033	46.2	2,508	19.2	2,041	15.6	562	4.3	197	1.5	13,052
Air Force	10,710	20.0	25,706	48.0	8,370	15.6	6,292	11.8	1,750	3.3	720	1.3	53,548
Total DoD	34,414	20.4	78,925	46.7	25,225	14.9	20,099	11.9	6,588	3.9	3,709	2.2	168,960

Source: Defense Manpower Data Center.

*Aptitude categories are based on scores of individuals who took the SAT between 1972 and 1981. The percentiles comprising the categories are as follows: O-I = 93-100; O-II = 65-92; O-IIIA = 50-64; O-IIIB = 31-49; O-IV = 10-30; and O-V = 1-9.

Officer Performance Measures in Relation to Aptitude

Given this cohort of officers with aptitude test scores, a rare opportunity exists to study officer performance in relation to aptitude across all four Services. Certain constraints were operating in relation to the particular database and in studying officers in general which are not usually present in analyses of enlisted personnel. Some of the data fields that are included in DMDC's master file were not present in the SAT/officer file. This complicated attrition and retention analyses, as well as analyses by source of commission. Additionally, restriction of range is ever present in officer aptitude and performance measures. Since this is a profoundly select group, there is little variance. It can be likened to studying the variance in grade point averages of graduate students. Despite these constraints, three basic surrogate measures of job performance were used in these analyses: (a) promotion, (b) retention, and (c) attrition.

Promotion Data Analysis. Table 3 shows correlations between SAT scores and number of months to promotion to O-3, or Captain, by commissioning source. It should be noted that the only available categorization of commissioning source for the SAT/officer cohort was academy versus non-academy. Although this two-way break is by no means ideal, these categories can be useful.

Several correlations in Table 3 are statistically significant but practical significance is questionable in light of the large numbers of officers and small coefficients. There is a relatively stable tendency for the larger sample sizes to have significant results. What is lacking is any other logical or explainable trend. Significant correlations of .02, .03 and even .06 must be taken as inconclusive evidence of the relationship between aptitude and time to promotion.

Table 3
Correlations Between SAT Score and Number of Months to
Promotion to O-3 for FY 1975-81 Commissioned Officers¹

Service	Academy		Non-academy		Total	
	N	r	N	r	N	r
Army	4,523	0.03 *	12,330	0.01 ns	16,853	0.01 ns
Navy	3,811	0.00 ns	9,660	0.03 **	13,471	0.02 ns
Marine Corps	439	-0.23 **	3,545	-0.05 **	3,984	-0.06 **
Air Force	5,140	-0.01 ns	16,058	-0.04 **	21,198	0.02 **

Source: Defense Manpower Data Center.

¹ N includes only officers promoted to O-3.

ns not significant.

** significant at the .01 level.

* significant at the .05 level.

Retention Data Analysis. Table 4 shows the correlations between SAT score and number of months served. To analyze retention, active duty obligation is useful for determining whether individuals are staying in Service beyond their required time. Because there is no way to determine the active duty obligations with the present data file, correlations are shown comparing academy and non-academy, the only available categorization of commissioning source. Again, the issue of large Ns and small correlation coefficients crops up and leads one to question the significant correlations. No sizable or significant relationships were found for academy officers and the small significant correlations for non-academy officers were both negative. Rather than construe this as evidence of a true inverse relationship between

aptitude and retention, it is more likely the power of large sample sizes that produced significance. However, the negative trend may suggest that higher aptitude officers (who may be serving in civilian valued jobs, e.g., pilots) are somewhat more likely to depart from the military.

Table 4
Correlations Between SAT Score and Number of Months
Served for FY 1975 Commissioned Officers

Service	Academy		Non-academy		Total	
	N	r	N	r	N	r
Army	791	-0.02 ns	2,673	-0.10 **	3,464	-0.04 *
Navy	646	0.00 ns	1,739	-0.06 *	2,385	-0.04 *
Marine Corps	75	-0.06 ns	1,095	0.04 ns	1,170	0.05 ns
Air Force	747	0.00 ns	2,458	-0.01 ns	3,205	0.04 *

Source: Defense Manpower Data Center.

ns not significant.

** significant at the .01 level.

* significant at the .05 level.

Attrition Data Analysis. For all intents and purposes, using the present database, attrition is synonymous with the above definition of retention. Table 5 shows six year attrition rates by SAT category. There appears to be no consistent trend between attrition and aptitude level. Each Service varies as to which aptitude categories have the highest attrition rates. It is assumed that these data probably reflect varying Service policy. As with the retention data, a tendency for higher attrition to occur at the lower aptitude levels (as shown for the Navy, Marine Corps, and Air Force) could be a byproduct of commissioning sources that select on SAT scores and incur a longer service obligation. Since neither active duty obligation nor commissioning source were available, it is difficult draw meaningful interpretations of these data.

Table 5
Percent of FY 1975-79 Commissioned Officers Attrited
Within 6 Years by SAT Category*

Service	O-I		O-II		O-IIIA		O-IIIB		O-IV		O-V		Total
	N	%	N	%	N	%	N	%	N	%	N	%	N
Army	3,239	44.5	7,971	40.2	2,790	40.8	2,491	38.3	953	36.5	578	42.0	18,022
Navy	4,058	38.7	7,814	34.8	1,661	36.1	897	40.5	189	40.2	65	43.1	14,684
Marine Corps	652	45.7	2,419	41.9	1,077	42.9	862	46.4	251	52.2	103	47.6	5,364
Air Force	3,301	21.7	7,969	21.3	2,345	26.1	1,803	26.7	521	20.9	260	25.4	16,199

Source: Defense Manpower Data Center.

*Aptitude categories are based on scores of individuals who took the SAT between 1972 and 1981. The percentiles comprising the categories are as follows: O-I = 93-100; O-II = 65-92; O-IIIA = 50-64; O-IIIB = 31-49; O-IV = 10-30; and O-V = 1-9.

Conclusions

What is abundantly clear from these data is that the officer corps is a highly select group. It is apparent that the officer selection system is working. The vast majority of officers in the SAT/officer cohort scored high on the SAT, with 67 percent categorized in this study within the high aptitude range. The relatively small proportion (six percent) of officers categorized as "low" aptitude is easily explained by several points. First, these SAT category IV and V are the lowest scoring in a very select group. Second, not all officers are screened on the SAT and not all colleges and universities screen on the SAT so it can be expected that there are college graduates who scored relatively low on the SAT. Additionally, when the SAT is used for screening, rigid cut-off scores are not typically applied.

What follows from such a selected group as officers is range restriction, which complicates the task of analyzing performance measures in relation to aptitude. Range restriction in the "predictor" (i.e., SAT scores) and in certain criteria (i.e., promotion), complicates such analyses, as was discovered. Although meaningful significant data were generally not found, one would not conclude that no relationship exists between aptitude and officer performance for several reasons. First, the job performance measures used here (promotion, retention, and attrition) are really surrogate measures. Supervisory or commander ratings of an individual may be more appropriately termed a job performance measure, albeit fraught with problems of range restriction. Second, the SAT is not designed to predict any such job performance measures. It is validated against first year college grade point average and is generally effective in predicting this criterion. The SAT provided the measure of aptitude for the present study because of its relatively consistent use in officer and college selection. Third, if validity generalization holds true, there probably is a relationship there for officers. What can be concluded is that given the present data, this relationship between aptitude and officer performance cannot be measured in a meaningful and significant way due to range restriction as well as constraints on the available data.

Although clear relationships between aptitude and officer performance were not found in the present study, this relationship has been demonstrated for the SAT as well as for other officer selection tests. It has been demonstrated that the officer corps is a highly select group in terms of aptitude, so cognitive ability is unquestionably present. Given this relationship, it can still be expected that factors other than aptitude operate in predicting who will be good officers. These factors may include motivation, leadership ability, college major, or career goals. And in fact, the Services screen for such attributes in addition to aptitude. Aptitude tests are used in conjunction with other variables which are empirically validated to determine how much weight various selection criteria should receive.

Present analyses should not be overlooked entirely for lack of clear ties between aptitude and performance. What can most be gleaned from these data is the measure of quality of the officer corps. This opportunity to view officers across Services on a single measure of aptitude has shown them to be an impressive group in terms of general aptitude. Furthermore, it is clear that high aptitude officers are serving the same amount of time as the lower aptitude officers. In fact, these data suggest that the Services are getting a good return on their investment in officer personnel, suggesting that there is little evidence in favor of the officer corps becoming more selective.

DEVELOPMENT OF A NAVAL OFFICER SELECTION TEST

Lieutenant (N) Alan C. Okros

Canadian Forces Personnel Applied Research Unit
Willowdale, Ontario, Canada

INTRODUCTION

The Naval Officer Selection Board (NOSB) employs an assessment centre approach in the evaluation of naval officer candidates (Rodgers, 1985). Ongoing Canadian Forces Personnel Applied Research Unit (CFPARU) work has sought to improve the NOSB's prediction of naval officer classification training performance. Based on an ability analysis (Rodgers, 1986) and a review of existing selection measures, a requirement was identified to develop a measure or measures of memorization and selective attention. This paper presents the development and evaluation of the test produced.

DEVELOPMENT OF THE PASSAGE PLANNING TEST

A review of existing commercial and Canadian Forces measures did not produce a suitable test. In this search it was, however, discovered that the U.S. Army Research Institute (ARI) was in the process of developing a test battery designed to assess the abilities required for U.S. Army helicopter pilot training. Included in this battery was a complex cognitive-perceptual test labelled the Flight Planning Test (FPT) which measured memorization, selective attention and decision-making. McAnulty, Cross and Jones (1986) reported that the FPT provided reliable measures of the three abilities while gain scores within levels produced unique assessments of two dynamic factors: learning and fatigue. Consequently, the FPT was obtained through The Technical Co-operation Panel (TTCP) of which both CFPARU and ARI are members.

The FPT was adapted for use as a potential naval officer selection measure through minor changes in the format, length and time allotted. The adapted test was labelled the Passage Planning Test (PPT). Using a paper-and-pencil, multiple-choice format, the PPT contains five timed sections presented in three levels of difficulty. Information is presented by a combination of words, numbers and symbols with some memorization of values and symbols required. Passage planning rules must be applied in a hierarchical order which becomes more complex in Level II and again in Level III. Selective attention is measured by including distracting information in one-third of the questions. The test is presented in booklet format containing all relevant information and can be self-administered. Time limits are provided for practice questions, the review of rules and the test sections. Each section contains 12 four-response multiple-choice questions, and responses are made on a separate machine readable form.

The views and opinions expressed in this paper are those of the author and not necessarily those of the Department of National Defence.

EVALUATION OF THE PASSAGE PLANNING TEST

Evaluation of the PPT was designed to answer three questions:

- a. were the psychometric characteristics of the PPT similar to those of the FPT?;
- b. did the PPT make a unique contribution to NOSB assessment?; and,
- c. did the PPT correlate with training performance?

The test was written by two groups: 61 civilian applicants attending an NOSB and 122 naval officers in training. Additional data collected included all existing selection measures and a measure of training performance. The analyses incorporated three different groupings of data. First, the entire sample was analyzed to compare the psychometric characteristics of the PPT to those reported for the FPT. Second, using the sub-group which attended an NOSB, the contribution of the PPT to the present assessment measures was examined. Finally, using the sub-group for whom training results were available, the concurrent validity of the PPT was established.

ANALYSES AND RESULTS

Psychometric Characteristics

The analyses revealed that, when time reductions were accounted for, the PPT exhibited similar characteristics and patterns of scores as had the FPT. As displayed in Table 1, mean scores generally decreased across the three levels and increased within the three sections of Level II. In comparison to the FPT, it was predicted that, due to the changes in length and times, the mean item difficulty would increase from .41 to .50 and the Level I mean score would decrease while the Level/III mean

Table 1

Descriptive Statistics for the FPT & PPT

Level	Section Order	FPT 1				PPT			
		Mean	SD	Time (Min)	Items Attempted	Mean	SD	Time (Min)	Items Attempted
I	1	6.28	2.26	9	11.2	6.12	2.10	6	9.75
II	2	4.70	2.25	9	8.5-8.9 ²	6.30	2.22	8	9.23
II	3	5.18	2.38	9	8.5-8.9 ²	6.73	2.10	8	9.99
II	4	5.84	2.64	9	8.5-8.9 ²	7.33	2.43	8	10.39
III	5	4.41	2.26	9	8.5-8.9 ²	5.46	2.05	10	9.79
N		273				183			

- Notes: 1. The FPT contains two additional, more difficult sections.
 2. Not reported by section.

score would increase. These predicted effects are evident in the reduction in the mean number of items attempted in section 1 (9.75 vice 11.2); however, the number attempted was generally greater for the remaining sections despite the reduced time on the Level II sections. The net gain scores in Level II (computed as the difference in score between section 4 and section 2) were highly similar (1.14 vice 1.03). As predicted, the mean item difficulty increased to .53, extremely close to the optimum level of .50. As reported for the PPT, section-total correlations were significant, correlations among the sections were moderate and gain scores were independent of total scores.

Contribution to Existing Assessment Measures

As reported in Johnston, Okros & Rodgers (1988), the NOSB merit score is computed from seven measures: a file review of background information, an interview, a conducting officer's assessment, a practical leadership exercise, a junior-manager-level in-basket exercise and leaderless group discussions (LGDs). NOSB candidates are pre-screened on the basis of two recruiting measures: a test of general learning ability (GC) and an assessment of military potential (MP). To be useful predictors, the PPT Total Score and Gain Score should generally have low correlations with these existing measures (some inter-measure correlations are expected in any selection battery). As displayed in Table 2, correlations for the PPT Total Score are generally non-significant or low. The one moderate correlation (GC: $r = .44$) was expected as both are cognitive measures. The PPT Gain Score did not produce any positive correlations indicating that the dynamic learning ability represented is not measured by any NOSB or recruiting measure.

Table 2
PPT Total and Gain Score Correlations
With NOSB Measures

Existing Measures	PPT Measures	
	Total Score	Gain Score
NOSB Measures		
Conducting Officer	--	--
Interview	.16	-.16
Leadership Assessment	.26	--
Leaderless Group Discussion 1	--	-.19
Leaderless Group Discussion 2	--	--
File Review	.22	--
In-Basket Exercise	--	--
NOSB Merit Score	.22	-.18
Recruiting Measures		
GC Test	.44	--
Military Potential	--	--

Concurrent Validity

As shown in Table 3, the PPT Total Score correlated significantly with the training performance measure ($r = .32$). Significant but lower correlations with the training measure were also found for the GC ($r = .21$) and the NOSB Merit Score ($r = .17$). The PPT Gain Score was not correlated with the training performance measure.

Table 3

PPT, GC and Merit Score Correlations
With Naval Officer Training Performance

Predictor	MARS III
PPT Total Score	.32
PPT Gain Score	--
GC Test	.21
NOSB Merit Score	.17

N=64

DISCUSSION AND RECOMMENDATIONS

The research was conducted to evaluate the psychometric properties and concurrent validity of the PPT. Comparison to the FPT revealed a highly similar pattern of results. Generally, scores declined as difficulty increased; however, scores increased within Level II. Analyses of item difficulties and item-total, section-total and gain score-total correlations demonstrated that the PPT possesses desirable psychometric characteristics.

The results indicated that the PPT made a unique contribution to the NOSB assessment. An expected moderate correlation was found between the PPT Total Score and the GC. Of interest, the dynamic learning ability tapped by the Gain Score was not assessed by any of the existing NOSB or recruiting measures. The PPT was also found to be a valid predictor of naval officer training performance. Both the GC and the NOSB Merit Score were also found to correlate with training performance. Although the sample size was small, the PPT correlation was greater than that obtained for the GC; thus, although the measures are correlated, the PPT is not simply a duplicate measure of general learning ability. As presented by McAnulty et al. (1986), a complex measure which permits the demonstration of learning and minimizes the effects of prior experience will produce a better index of a dynamic process like naval officer classification training than will a static measure. Based on these encouraging results, the PPT has been incorporated at the NOSB as a selection test.

As a concluding comment, this project has served to illustrate the benefits which may be derived from cooperative international forums such

as MTA and TTCP. Clearly the credit for designing, refining and pre-testing the PPT belongs to ARI and the savings in time and effort which accrue from shared work are substantial. These results have been forwarded to ARI as will those from future work conducted with the PPT. It is hoped that this continued exchange of information will prove fruitful in the future.

REFERENCES

- McAnulty, D.M., Cross, K.D., & Jones, D.H. (1986). The development of an experimental battery of aviation-related ability tests. (Draft Technical Report). Fort Rucker, Alabama: United States Army Research Institute Aviation Research and Development Activity.
- Okros, A.C., Johnston, V.W., & Rodgers, M.N. (1988). An evaluation of the effectiveness of the Naval Officer Selection Board as a predictor of success on the Basic Officer Selection Course (Working Paper 88-1). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.
- Rodgers, M.N. (1985, June). Assessment centre approach to officer selection in the Canadian Forces. Paper presented at the 21st International Applied Military Psychology Symposium, Paris, France.
- Rodgers, M.N. (1986). The identification of ability requirements for MARS officer selection (Technical Note 23/86). Willowdale, Ontario: Canadian Forces Personnel Applied Research Unit.

COMPLEX COGNITIVE ASSESSMENT BATTERY:
PERFORMANCE, DEMOGRAPHIC AND ATTITUDINAL ASPECTS

William D. Sprenger, Ph.D.
U.S. Army Research Institute (ARI)
and
Shippensburg University

Jon J. Fallesen, Ph.D.
U.S. Army Research Institute
Ft. Leavenworth Field Unit, KS

Background

The Complex Cognitive Assessment Battery (CCAB) was developed to assess higher-order cognitive capabilities as is representative of problem solving and decision making abilities in military command and control tasks (see the appendix for level of association between CCAB constructs and subtests). The primary testing application for which the CCAB was developed originally was the determination of the effects of environmental stressors on complex cognitive performance. The CCAB underwent an extensive and thorough conceptual development phase, which began with the development of constructs for complex cognitive capabilities and tests were then selected or developed tests to meet a number of criteria (for a more complete description, see Samet and Hartel, 1988). The constructs are based on an information processing model consisting of four categories: (1) responding to data, (2) going beyond data, (3) taking action based on data, and (4) creating data. Specific capabilities or functions were assigned under each category, e.g. category 3 is comprised of planning, situation assessment and decision making (Samet, 1987).

In the configuration used in the present study, the CCAB consists of a battery of nine computerized tests which address all capabilities from all four constructs. The battery of tests is administered using a PC-AT or PC-XT computer with 384K of memory, hard disk and graphics card. Each test consists of a set of instructions, practice problems, a sample problem to evaluate the subject's understanding of the task, and a set of test problems. The CCAB automatically performs all necessary scoring and summary actions. The battery can be configured under the tester's control in terms of which tests are used, the sequence and the number of repetitions for each test.

Description of CCAB Sub-tests:

1. Tower Puzzle (TP)--This test is the classic Tower of Hanoi puzzle which requires the movement of five disks, each of a different size, among three posts. The objective is to move all the disks to one post so that no larger disk is ever above a smaller one.
2. Following Directions (FD)--Subjects must follow directions to move a cursor among several lines of text and take the indicated actions to mark or underline words in the text.

3. Word Anagrams (WA)--In this variation of the classic anagram problem, the subject is required to make as many as three or more letter words from two unique vowel and four unique consonants.

4. Logical Relations (LR)--The participant must examine two premises depicting relations, such as worse than/better than or longer than/shorter than, and must reason which item has the greatest or smallest value (logical syllogism).

5. Mark Numbers (MN)--This test consists of two simultaneous tasks, one requires the subject to search an array of numbers to see if it fits some specified criteria, such as mark odd numbers between 10 and 28, the other task requires the marking of one of two numbers which are flashing in the array.

6. Numbers and Words (NW)--In this simultaneous task, previously displayed numbers must be recalled each time a new one is displayed, and a word must be distinguished as it gradually appears from an indiscernible scattering of dots.

7. Information Purchase (IP)--A table containing 14 columns and 6 rows is displayed. Initially only the first column is filled from the digits of 0, 1, or 2. The subject can ask for the next column to be filled in and the next and so on, until an estimate is made about which row will have the largest sum.

8. Route Planning (RP)--A 5 x 5 matrix is presented with 11 squares randomly shaded. The subject must plan a route through the matrix such that a shaded square is never landed on.

9. Missing Items (MI)--The subject must determine which item is missing from a sequence of items with a logical order or pattern of numbers or letters. (Samet, Geiselman, Zajackowski, and Marshall-Mies, 1987)

The scoring techniques for each test differ but rely on such measures as total time, correct response, number of errors, reaction time, and time between responses. The CCAB software makes available a 15 page scoring report for a participant.

Purpose of Research

This research was conducted in order to field test the CCAB to determine the intercorrelations of the subtests, descriptive statistical qualities, efficacy of subtest instruction and the suitability of the CCAB for possible future use with the Command and Control Performance Assessment (C2PAS) System which is undergoing development at the Ft. Leavenworth Field Unit.

Subjects

Fifty-one subjects volunteered for this research study, all from Shippensburg University courses taught by the senior author. They were enrolled in General or Educational Psychology courses and they were given the option of taking the CCAB or performing another similar assignment.

Data Collection Methods Used

Each student was scheduled for a two hour block of time to take the CCAB in an office. The test administrator explained the procedures of the test, provided them with a privacy act statement and left the room shortly after each subject began the test. After completion of the CCAB, a survey instrument was filled out to determine the attitude toward the CCAB and testing conditions. The quickest subject completed the test in 81 minutes and the slowest took 122 minutes.

Results

Table 1 lists the mean and standard deviations of the nine subtests and includes the composite score of the CCAB (TMEAN) and selected demographic variables. The composite mean score was 1008 with a standard deviation of 120 (lowest score was 671; highest was 1266). The least difficult subtest was the missing items subtest and the most difficult was the tower puzzle; however, since the tower puzzle (in this study) was always presented first, at least part of the low score may be a function of learning to take computer generated tests.

The average age of the participants was 20.3 years with gender being about evenly divided. The average total SAT score was 968 and the average GPA was 2.5 which is typical of most colleges and is considered appropriate for engaging in university level academic endeavors. The participants found the test to be about average in difficulty (2.5 on a scale of 1 to 5). They also found the test instructions to be adequate. Most participants reported low levels of personal stress during the time that they took the CCAB (average score was 2.1 on a score of 1 to 5).

TABLE 1
MEANS AND STANDARD DEVIATIONS DEMOGRAPHIC VARIABLES
AND CCAB SUBTESTS (N=51)

CCAB	MEAN	S.D.	DEMOGRAPHICS	MEAN	S.D.
TOTAL MEAN (TMEAN)	1008	120	AGE	20.3	6.0
TOWER PUZZLE (TP)	782	391	COLLEGE CREDITS	20.7	15.4
FOLLOWING DIRECTIONS (FD)	1075	159	GPA	2.5	.6
WORD ANAGRAMS (WA)	1013	151	SAT-VERBAL	466.6	62.4
LOGICAL RELATIONS (LR)	949	250	SAT-MATH (M)	498.5	71.6
MARK NUMBERS (MN)	1112	210	TOTAL-SAT (TSAT)	968.7	92.3
NUMBERS AND WORDS (NW)	971	206	DIFFICULTY	2.7	.6
INFORMATION PURCHASE (IP)	972	93	INSTRUCTIONS	1.1	.3
ROUTE PLANNING (RP)	963	169	STRESS	2.1	.1
MISSING ITEMS (MI)	1233	151			

Table 2 shows Pearson correlations of the nine CCAB subtests with selected demographic variables. The word anagram subtest shows statistical significance ($p < .01$) with the SAT-Verbal and Total SAT (Math and Verbal) score as does Logical Relations with the SAT-Math and the Total SAT Score. The Word Anagram subtest also shows statistical significance ($p < .05$) with

age, college credits and SAT-Math. The Mark Numbers subtest shows a significant correlation ($p < .01$) with SAT- Math and also at the .05 level with the total SAT score. The correlations suggest that those developed or fluid abilities necessary to score well on the SAT test are also necessary to score well on the CCAB.

TABLE 2
PEARSON CORRELATION COEFFICIENTS
CCAB AND SELECTED DEMOGRAPHIC VARIABLES (N=51)

DEMOGRAPHIC VARIABLES	TP	FD	WA	LR	MN	NW	IP	RP	MI
AGE	-.06	-.27*	-.25*	-.33*	-.16	-.24	.04	.18	-.00
COL.CRDTS.	-.01	-.21	-.23*	-.23*	-.18	-.08	.06	.08	-.10
GPA	-.11	-.24*	-.27*	-.21	-.34	-.14	-.04	-.24*	-.05
SAT-V	-.12	-.33	.32**	.15	-.12	.35	-.15	.20	-.05
SAT-M	.28*	-.38*	.29*	.44**	.46**	.24	-.17	.01	.07
SAT-TOTAL	-.29*	.51*	.50**	.48**	.28*	.46	-.21	.17	.21
DIFFICULTY	-.18	-.07	.05	-.32**	-.16	-.16	.10	.21	.26
INSTRUCT.	-.04	-.14	-.10	-.22	-.09	-.08	.26	.15	-.11
STRESS	-.12	-.24	-.19	.06	-.05	.08	.03	-.09	.00

* $p < .05$

** $p < .01$

Table 3 shows that shows 26 significant (and 19 nonsignificant correlations) either at the .01 or .05 level which demonstrates that a number of the subtests seem to measure similar cognitive abilities. Of the nine possible correlations with the total mean score (composite) of the CCAB eight show statistical significance at the .01 level which one would expect of an auto-correlation.

TABLE 3
INTERCORRELATIONS
PEARSON CORRELATION COEFFICIENTS
COMPLEX COGNITIVE APTITUDE BATTERY (CCAB). (N=51)

	TMEAN	TP	FD	WA	LR	MN	NW	IP	RP	MI
TMEAN	1.00									
TP	.67**	1.00								
FD	.59**	.22	1.00							
WA	.67*	.29*	.58**	1.00						
LR	.74*	.35**	.35**	.40**	1.00					
MN	.6**	.30*	.42**	.52**	.52**	1.00				
NW	.70**	.27*	.38**	.45**	.56**	.37**	1.00			
IP	.06	-.08	-.17	-.11	.05	-.11	.07	1.00		
RP	.37**	.00	.18	.20	.10	.14	.39**	.22	1.00	
MI	.50**	.38**	.21	.24*	.31*	.13	.14	.12	.12	1.00

* $p < .05$

* $p < .01$

Table 4 is an attempt to provide some tentative norming for the CCAB (as configured for this study) based on 51 participants at Shippensburg University. They were mostly freshman and sophomore students with average SAT Scores. A composite mean score of 600 on the CCAB would probably be an indication that a participant did not really take the test. The CCAB will produce a score at this level with no responses from the subject. One would have to artificially take the exam to produce a score above 1500, that is only an expert with prior knowledge of the correct response could attain such a score.

TABLE 4
TENTATIVE COMPOSITE SCORE NORMS FOR THE CCAB
BASED ON COLLEGE AGE POPULATION

SCORE LEVEL	
600---756	UNUSUALLY LOW
751---876	LOW
877---995	LOW AVERAGE
996--1127	AVERAGE
1128--1240	ABOVE AVERAGE
1241--1368	SUPERIOR
1369--1500	UNUSUALLY HIGH
1501+	ALMOST IMPOSSIBLE TO OBTAIN

Discussion and Conclusion

The CCAB test seems to measure similar cognitive functions as seen by the large number of significant intercorrelations and therefore in the future fewer subtests may be needed in assessing performance (and prediction) of staff officers in future research using the C2PAS and the Experimental Development, Demonstration and Integration Center "(EDDIC)" that are currently being developed at the ARI Ft. Leavenworth Field Unit. In addition, we now have descriptive data for each subtest that will be useful in planning future research on staff officer performance in tactical problem solving.

References

- Samet, M.G. and Hartel, C.R. (1988). High level cognitive skills: Complex Cognitive Assessment Battery. Proceedings of the 29th Annual Military Testing Association, 339-344.
- Samet, M.G. (1987). Complex Cognitive Assessment Battery (CCAB): Taxonomy Development. Research Report, Analytical Assessments Corporation, Los Angeles, CA.
- Samet, M.G., Geiselman, R.E., Zajackowski, F., and Marshall-Mies, J. (1987). Technical User Manual, Analytical Assessments Corporation, Los Angeles, CA. Complex Cognitive Assessment Battery (CCAB): Test Descriptions.

APPENDIX

LEVEL OF ASSOCIATION BETWEEN CCAB CONSTRUCTS AND TESTS* (1=low; 2=Medium; 3=High)

Cognitive Complexity Categories	Cognitive Construct Measures	CCAB TESTS**								
		TP	FD	WA	LR	MN	NW	IP	RP	MI
1. Responding to Data	Attention to Detail		3	1	1		2	2	1	2
	Perception of Form	1		2			3	1	2	2
	Memory Retrieval		2	3	1	1	2	1		
	Time Sharing		2			2	3	1		
2. Going Beyond Data	Comprehension		2		3	1				
	Concept Formation	1		1	1		2	1	1	3
	Verbal Reasoning		2		3	1				
	Quantitative Reasoning	1	1		2	3		2	1	3
3. Taking Action Based on Data	Planning	3		2				2	3	
	Situation Assessment	3				1		3	2	1
	Decision Making	2		2			2	3	1	
4. Creating Data or Solutions	Communication	1	1							
	Problem Solving	3		1	1			1	2	2
	Creativity	1		3						

*Source: CCAB User Guide

**CCAB consists of 9 tests. Codes used in the table for tests are as follows: Tower puzzle (TP), Following Directions (FD), Word Anagrams (WA), Logical Relations (LR), Mark Numbers (NM), Numbers and Words (NW), Information Purchase (IP), Route Planning (RP), and Missing Items (MI).

SUBTEST MEAN SCORES

Tower Puzzle	TP	782
Following Directions	FD	1075
Word Anagrams	WA	1013
Logical Reasoning	LR	949
Mark Numbers	MN	1112
Numbers and Words	NW	971
Information Purchase	IP	972
Route Planning	RP	963
Missing Items	MI	1233

APPLICATION OF PERFORMANCE PROTOCOL ANALYSIS IN MILITARY TESTING

Michael G. Samet
Integrated Systems Research Corporation
Woodland Hills, CA 91367

Christine Hartel
U.S. Army Research Institute
Alexandria, VA 22333

In recent years, the role of high-level cognitive testing in computer-based command and control environments has received much attention. In particular, it is becoming more important to evaluate the process used by an operator (i.e., the test subject) in making tactical judgments and decisions as opposed to measuring only the outcome of this process. That is, for many testing purposes, how a problem solution is arrived at can be just as significant as whether that solution is correct or not.

A testing methodology is described here that involves the on-line collection and analysis of all user actions (e.g., keypresses, text and graphic information selection and viewing, map utilization, time pacing, etc.) during the performance of a computer-aided tactical problem solving session. This methodology has been embedded into the C2PAS, a Command and Control Performance Assessment System whose development has been supported by ARI. This videodisc-based microcomputer system is designed to present simulated tactical scenarios, battle situations, and problems to military subjects so that their performance in solving these problems can be assessed.

A sample application of the embedded-performance testing methodology is described below. Sample protocol data are presented, including usage pattern analyses that can be used to derive normative representations of cognitive models of user interaction with the C2PAS system. Such user models can serve as performance standards or criteria by which the individual performance models of other operators can be evaluated.

SAMPLE APPLICATION

A pilot study of the Command and Control Performance Assessment System (C2PAS) was performed to test how staff officers would use the operational and functional features of this new system. Several participants with suitable military experience were introduced to the system, and then they were presented with selected tactical scenarios and sample problems. The participants used the features of the system to study the relevant tactical situation/map displays and available information to arrive at answers for the tactical problem. As the participants worked through each tactical problem, their step-by-step performance protocols were recorded by the system for subsequent examination.

Method Formal testing sessions of three hour duration were established for each of the test participants. Prior to beginning work, each participant was given an overview of the C2PAS system including its purpose

and general operational features. Then the participant was allowed to manipulate some of the user interface features and to ask any questions about system operation. This introductory procedure took only about 20 minutes.

Following this orientation, the participant began a practice session with a sample scenario that included typical tactical data and required the same problem solution approaches that made up the pilot-study test scenarios themselves. The participants were told to experiment with the features of the system, and to move at a comfortable pace through the practice session. For most participants, the practice session lasted between 20 and 30 minutes.

Once the participant completed the practice session, he was ready to begin work on the actual scenarios, vignettes, and tactical problems that were presented in the pilot study. Each tactical scenario (one for defense, and one for counter-offense) consisted of a series of vignettes, each prefaced by a set of Build-up text displays, and each consisting of its own tactical unit data (i.e., unit positions and associated unit information), and control measures (i.e., battlefield lines, borders, objectives, etc.). Within the context of a scenario and a vignette, specific tactical problems were presented to the participants for solution.

An example of the type of tactical problems use is: Assign axes of advance to lead teams. Also, which lead team would the reserve team follow? Each participant completed between three (3) and five (5) tactical problems, which were presented in the context of either one or both of the tactical scenarios. Each tactical problem was associated with only one vignette.

Results. The results are based on a detailed examination of the performance protocol reports that are automatically recorded during the C2PAS session.

Table 1 provides a sample performance protocol recorded for one participant's tactical problem-solving session. The columns in the table give the time of each action the participant made, the status of the video and text displays, the menu accessed and the selection/choice made, and the UTM coordinates associated with the cursor position on the graphics screen. Based on this sample protocol, note how the participant systematically took advantage of many C2PAS features including zooming and panning of map displays, distance estimation, selective display of different types of units, and access of a variety of unit information and symbol feature information.

Table 2 illustrates a sample result summary report or "vignette analysis" that was generated by the C2PAS program, based on the performance protocol data for the given participant and scenario/vignette represented in Table 1. The report covered in the table includes frequency tables and histograms for a variety of user actions taken in selectively accessing tactical information overlays and manipulating map backgrounds during the course of one tactical problem-solving session. These actions include: differential use of available program displays and menus; map panning and zooming behaviors; and, access time and frequency for differential unit information and symbol features.

Discussion. The performance protocols reveal that participants made extensive use of the information call-up features of the system. The participants frequently queried the system for unit identification data, especially in distinguishing between the units that were Headquarter or Non-Headquarter units. Additional queries they made regarded "readiness" status, "unit type," "strength," "direction of movement," and other basic tactical information. Participants also made great use of the zooming and panning features of the C2PAS system, demonstrating the requirement for constant updating and verification of position identification.

These types of individual performance protocols can be aggregated over a group of participants to obtain a simple cognitive model or normative "schema" for how good performers respond to specific tactical questions. The schemata analyses are particularly useful to describe the performance of the participants whose answers to the problems were judged by experts to be "good". Thus, a schema analysis can be conducted for each tactical question studied using the overlay selections of those participants who prepared the good answers to that question. For example, problem required the participants to recommend whether or not, as well as where, any of the friendly artillery batteries should be moved. To answer this tactical question, the majority of participants were concerned primarily with the following tactical information and corresponding overlays.

- (1) Current location of the friendly artillery batteries.
- (2) Tactical reach of the friendly artillery batteries.
- (3) Identification and major weapons of the friendly artillery batteries.
- (4) Current location of the friendly units that the artillery is supporting.
- (5) Location of enemy targets.
- (6) Threats of friendly artillery batteries.

Thus, the schema analysis of the protocols leading to the good answers to this problem indicated that participants who did well first built a picture of their own forces, and then they added information dimensions about the enemy. That is, friendly unit locations, their ranges, weapons, and identity were selected first, and enemy information was subsequently added to this configuration. In contrast, schema analysis performed on the protocols that lead to poor answers on the same problem revealed a different information processing style. The participants who provided poor answers vacillated between the selection of friendly and enemy information. Although the same overlays were ultimately called up during the problem solution.

Overall, in conclusion, the behavioral measurement methodology described in this paper has implications for military selection, training, and aiding, as well as for cognitive test battery validation.

TABLE 1
Portion of Sample C2PAS Performance Protocol

SCENARIO RECORDING
VIGNETTE #1
PARTICIPANT: 2323

<u>TIME</u>	<u>VIDEO DISPLAY</u>	<u>TEXT DISPLAY</u>	<u>MENU</u>	<u>SELECTION</u>	<u>CHOICE</u>	<u>UTM</u>
000000		BUILDUP	TEXT	START EXERC		32UNB449289
000108	SIT MAP	PROBLEM	MAP	SIT MAP		32UNB449289
000112	SIT MAP	PROBLEM	DISPLAY	UNIT		32UNB449289
000115	SIT MAP	PROBLEM	TYPE	FR MECH INF HQ	ON	32UNB449289
000142	SIT MAP	PROBLEM	TYPE	FR MECH INF Non HQ	ON	32UNB479224
000162	SIT MAP	PROBLEM	TYPE	FR MECH INF Non HQ	OFF	32UNB479224
000165	SIT MAP	PROBLEM	TYPE	FR ARMORED HQ	ON	32UNB479224
000167	SIT MAP	PROBLEM	TYPE	FR ARMORED Non HQ	ON	32UNB479224
000169	SIT MAP	PROBLEM	TYPE	FR MECH INF HQ	OFF	32UNB479224
000204	SIT MAP	PROBLEM	DISPLAY	UNIT		32UNB479224
000216	SIT MAP	PROBLEM	CTRL MEAS	POLYGONS	ON	32UNB479224
000250	SIT MAP	PROBLEM	CTRL MEAS	LINES	ON	32UNB479224
000271	SIT MAP	PROBLEM	DISPLAY	UNIT		32UNB505198
000282	SIT MAP	PROBLEM	INFORMATION	BASIC	ON	32UNB505198
000296	SIT MAP	PROBLEM	INFORMATION	READINESS	ON	32UNB505198
000331	SIT MAP	PROBLEM	SYMBOL FEAT	DIR OF MOVEMENT	ON	32UNB505198
000356	SIT MAP	PROBLEM	SYMBOL FEAT	UNIT NAME	ON	32UNB505201
000397	SIT MAP	PROBLEM	DISPLAY	UNIT		32UNB505201
000411	SIT MAP	PROBLEM	INFORMATION	READINESS	ON	32UNB505201
000424	SIT MAP	PROBLEM	INFORMATION	BASIC	ON	32UNB505201
000431	SIT MAP	PROBLEM	INFORMATION	VEHICLES	ON	32UNB505201
000434	SIT MAP	PROBLEM	INFORMATION	STRENGTH	ON	32UNB505201
000601	SIT MAP	PROBLEM	INFORMATION	COMMO	ON	32UNB505201
000605	SIT MAP	PROBLEM	INFORMATION	WEAPONS	ON	32UNB505201
000693	SIT MAP	PROBLEM	DISPLAY	UNIT		32UNB505201
000716	SIT MAP	PROBLEM	SYMBOL FEAT	AVAILABILITY	ON	32UNB505201
000744	SIT MAP	PROBLEM	SYMBOL FEAT	AVAILABILITY	OFF	32UNB505201
000764	SIT MAP	PROBLEM	DISPLAY	UNIT		32UNB505201
000784	SIT MAP	PROBLEM	DISPLAY	UNIT		32UNB505201
000793	SIT MAP	PROBLEM	DISPLAY	UNIT		32UNB505201
000678	SIT MAP	PROBLEM	PAN	SOUTH		32UNB394309
000901	SIT MAP	PROBLEM	ZOOM	IN		32UNB394209
000916	SIT MAP	PROBLEM	PAN	EAST		32UNB394209
000933	SIT MAP	PROBLEM	PAN	SOUTH		32UNB434208
000940	SIT MAP	PROBLEM	ZOOM	OUT		32UNB434178
000971	SIT MAP	PROBLEM	ZOOM	IN		32UNB434177
000978	SIT MAP	PROBLEM	PAN	EAST		32UNB433177
000988	SIT MAP	PROBLEM	PAN	EAST		32UNB473177
001001	SIT MAP	PROBLEM	PAN	SOUTH		32UNB513176
001011	SIT MAP	PROBLEM	PAN	NORTH		32UNB513146
001022	SIT MAP	PROBLEM	PAN	NORTH		32UNB513176
001062	SIT MAP	PROBLEM	DISPLAY	UNIT		32UNB513207
001067	SIT MAP	PROBLEM	TYPE	FR MECH INF HQ	ON	32UNB513207

Table 2
Portion of Sample C2PAS Performance Analysis

elapsed time = 00:58:40

MENU	Frequency	Time	
MAP	1	003.06	**
DISPLAY	13	017.07	*****
LOCATION	2	002.78	**
PAN	10	016.61	*****
ZOOM	6	014.09	*****
RESTART	0	000.00	
TEXT	2	026.61	*****
Total		100.00	

ACTION	Frequency	Repetitions
PAN NORTH	3	1
PAN SOUTH	3	0
PAN EAST	3	1
PAN WEST	0	0
PAN NORTHEAST	0	0
PAN NORTHWEST	1	0
PAN SOUTHEAST	0	0
PAN SOUTHWEST	0	0
ZOOM IN	3	0
ZOOM OUT	3	0

INFORMATION	Frequency	Time
BASIC	3	000.99
WEAPONS	4	000.93
VEHICLES	2	001.02
STRENGTH	2	000.39
SURVEILLANCE	0	000.00
NBC	0	000.00
READINESS	4	001.30
DECEPTION	0	000.00
COMMO	1	004.74

Table 2
Portion of Sample C2PAS Performance Analysis (continued)

```

FR MECH INF HQ      : **                                     *****
FR ARTILLERY HQ     : |
FR ARMORED HQ       : | *****
FR SCOUT HQ         : |
FR OBSTACLE HQ      : | *****
FR AIR MOBIL HQ     : |
FR FGHT BOMB HQ     : |
FR OTHERS HQ        : |
FR MECH INF Non HQ  : *               *****
FR ARTILLERY Non HQ : |
FR ARMORED Non HQ   : | *****
FR SCOUT Non HQ     : |
FR OBSTACLE Non HQ  : |
FR AIR MOBIL Non HQ : |
FR FGHT BOMB Non HQ : |
FR OTHERS Non HQ    : |
EN MECH INF HQ      : | *****
EN ARTILLERY HQ     : | *****
EN ARMORED HQ       : | *****
EN SCOUT HQ         : | *****
EN OBSTACLE HQ      : | *****
EN AIR MOBIL HQ     : |
EN FGHT BOMB HQ     : |
EN OTHERS HQ        : |
EN MECH INF Non HQ  : | *****
EN ARTILLERY Non HQ : | *****
EN ARMORED Non HQ   : | *****
EN SCOUT Non HQ     : | *****
EN OBSTACLE Non HQ  : | *****
EN AIR MOBIL Non HQ : |
EN FGHT BOMB Non HQ : |
EN OTHERS Non HQ    : |

```

-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

:	:	:	:	:	:	:	:	:	:
09:00:00	00:06:44	00:13:29	00:20:13	00:26:58	00:33:42	00:40:27	00:47:12	00:53:56	

Predicting Tank Gunnery Performance from Crewmembers' Experience and Cognitive Ability¹

R. Gene Hoffman
David A. Campshure
Human Resources Research Organization

A major portion of tank gunnery training consists of a series of exercises prescribed by FM 17-12-1, Tank Combat Tables (1986). For the individual tank, Table VIII is the culminating exercise and is used to evaluate tank crew gunnery proficiency. This report investigates two individual level variables that may be predictive of tank crew performance on Table VIII. The first is crew member experience and the second is cognitive ability as measured by the Armed Services Vocational Aptitude Battery (ASVAB). It is based on performance data for 872 M1 tank crews from 19 M1 tank battalions who fired the Table VIII range in Grafenwöhr, West Germany.

Performance Variables

Table VIII requires tank crews to fire main gun and COAX engagements consisting of one or two targets each, including both day and night engagements. Each crew fires 10 engagements, but not all crews fire the same 10.

For each crew, six variables were obtained or constructed to describe overall performance on Table VIII. These include (1) average points, (2) average cuts for procedural errors, (3) average score, (4) proportion of first round hits (excluding COAX engagements), (5) total proportion of hits, and (6) average opening time. Scoring of points, score and cuts is specified by FM 17-12-1. Although it is more common to use total scores, average scores were required to accommodate the decision to drop low frequency engagements from the analysis. Average score provides a composite picture of performance and was considered the primary variable for this analysis. It is a function of time, hits and procedural errors.

Predictor Variables

Two of the ASVAB composite scales were used in the analysis. The AFQT composite is the basic entry selection scale used Army wide. The Combat (CO) composite, formed from the arithmetic reasoning, coding speed, automotive/shop information, and mechanical comprehension subtests, is used to select recruits into Armor. Time in position was used as the index of experience. Time in position was available for all crew members. ASVAB scores were obtained for only a portion of the crew members.

Battalion Differences and Potential Criterion Contamination

Preliminary investigation showed significant mean differences in crew criterion scores across the 19 battalions. These battalion differences are likely to be a result of a variety of influences. That is, battalion

¹This research is funded by the Army Research Institute Contract No. MDA 903-86-C-0335. All statements in this paper are those of the author and do not necessarily express the official opinions of the U.S. Army Research Institute or the Department of the Army.

membership itself would not be considered a causal variable, but a classification index confounded with one or more causal variables. Some influences represented by the battalion differences would be considered extraneous to the present analysis and therefore represent contamination in the crews' criterion scores. For example, weather conditions and differences in target arrays are confounded with battalion and may affect crew performance. Statistical control of such influences would be advisable in order to allow the examination of individual variables and crew differences. On the other hand, to the extent that battalion differences are associated with differences in crew members' ability or experience, then the battalion differences do not represent criterion contamination. Instead, they represent the cumulative effects of the variables we are examining. The "truth" may be somewhere in between with part of the battalion differences due to criterion contamination and part due to the effects of the variables we are studying. Because of this, the following analysis of experience and cognitive ability were conducted with crews' battalion membership as an additional variable.

Results and Discussion

Correlations. Table 1 presents correlations between the predictor and criterion variables. Although the effects are not strong, both tank commander's and gunner's time in position are significantly related to average score, average cuts, and average opening time. Gunner's ASVAB Combat score is related to average score, average cuts and total hit probability. Gunner's AFQT is related to average cuts. No other positions show ability scores related to crew performance. Correlations for tank commander's AFQT with average score and total hit probability have values nearly equal to the gunner's correlations for Combat score with these same variables. However, the sample sizes are smaller and the correlations cannot be judged different from zero. Battalion differences were also found for time in position for both tank commander and gunner.

Table 1

Correlations Between Criterion and Predictor Measures

	n	Average Score	Average Cuts	Average Opening Time	Total Hit Probability	First Hit Probability
Tank Commander Time in Position	872	0.10*	-0.07*	-0.10*	0.05	0.02
Gunner Time in Position	872	0.09*	-0.14*	-0.11*	0.05	0.02
Loader Time in Position	872	0.01	0.00	-0.04	0.01	-0.01
Driver Time in Position	872	0.03	-0.02	-0.03	-0.02	0.00
Tank Commander AFQT Score	123	0.13	-0.10	0.07	0.12	0.11
Tank Commander Combat Score	117	0.02	-0.01	0.12	0.05	0.11
Gunner AFQT Score	395	0.04	-0.13*	0.05	0.08	0.04
Gunner Combat Score	392	0.13*	-0.10*	-0.03	0.12*	0.04
Loader AFQT Score	379	-0.04	-0.01	-0.02	-0.05	-0.09
Loader Combat Score	378	-0.07	0.06	0.02	-0.06	-0.07
Driver AFQT Score	357	0.02	-0.04	-0.01	0.01	-0.01
Driver Combat Score	357	0.01	-0.01	0.02	0.04	0.00

* p < .05

Composite equations. The major objective of the analysis was to build multiple regression equations showing the combined effects of battalion membership, crew members' experience and/or crew members' ability on those criterion variables demonstrated to have simple zero-order relationships with ability or experience.

For each criterion, a step-down multiple regression process was used where all variables with significant zero order correlations were entered in the equation. Then, one at a time beginning with the smallest contributor, variables were eliminated if their unique contribution to the equation was not statistically significant. This eliminated predictors whose covariance with the criterion was shared with, and accounted for by, other predictors in the equation.

Average score had significant zero-order correlations with tank commander's time in position, gunner's time in position, gunner's ASVAB Combat score, as well as battalion membership. In addition, curvilinear effects for gunner's time in position and tank commander's time in position were found for average score. These variables and the time in position curvilinear terms were entered into a regression equation. The curvilinear term for tank commander's time in position did not contribute significantly to the regression solution; the remaining variables each contributed significantly ($p < .05$) and were retained.

The solution for predicting average score is presented in Table 2. Significance levels for each variable were computed with the remaining variables entered in the equation. Thus, they represent the strength of the unique contribution each made to the prediction. Unique variance accounted for by each variable is shown in the last column, with the effects of time in position summed over the linear and quadratic terms. The table also shows the sum of the unique variance-accounted-for estimates. The difference between this value (21.8) and the total variance accounted for ($R^2 \times 100 = 22.9$) indicates the amount of predictable variance in average Table VIII scores that could not be uniquely assigned to any one variable. That is, the variables share in accounting for 1.1 percent of the variance in Table VIII score.

Table 2

Composite Effects of Battalion, Tank Commander Experience, Gunner Experience, and Gunner ASVAB Combat Score on Table VIII Average Score

Source	Sum of Squares	DF	Mean Square	F-Ratio	p	Coefficient	Unique Variance
Battalion	7103.35	18	394.63	4.407	.001	----	16.8%
Tank Commander Time in Position	347.56	1	347.56	3.881	.05	.122	.8%
Gunner Time in Position	721.19	1	721.19	8.054	.01	.437	2.9%
Gunner Time in Position Squared	497.71	1	497.71	5.560	.02	-.010	
Gunner Combat Score	549.92	1	549.92	6.141	.02	.084	1.3%
Error	33043.42	369	89.56				Total: 21.8%
R = .478 R ² = .229 p < .01							

Table 2 also presents regression coefficients for the variables, with the exception of battalion membership². Each regression coefficient indicates the expected change in average Table VII score given a unit of change in the predictor score, other things being equal.

Table 3 presents the regression results for average Table VIII cuts. The analysis of average cuts began with battalion membership, tank commander's time in position (linear and quadratic terms), gunner's time in position (linear and quadratic terms), gunner's AFQT and gunner's Combat score. In a step down solution, battalion membership, gunner's time in position (both terms) and gunner's AFQT are the only variables to remain in the equation. In order to check the relationship of gunner's Combat score independently of gunner's AFQT, the analysis was also conducted without gunner's AFQT. After battalion and experience variables are accounted for, gunner's Combat score is not significantly related to the criterion.

Table 3

Composite Effects of Battalion, Gunner Experience, and AFQT Percentile Score on Table VII Average Cuts

Source	Sum of Squares	DF	Mean Square	F-Ratio	p	Coefficient	Unique Variance
Battalion	109.93	18	6.11	2.85	.001	---	11.5%
Gunner Time in Position	25.13	1	25.13	11.74	.01	-.08	3.5%
Gunner Time in Position Squared	8.62	1	8.62	4.03	.05	.001	
Gunner AFQT Score	10.75	1	10.75	5.02	.05	-.007	1.1%
Error	798.51	373	2.14				Total: 16.1%
				R = .433	R ² = .187	p < .01	

Figure 1 illustrates the curvilinear effects of gunner's time in position on average score and average cuts. The curve for average cuts was constructed from the regression analysis presented in Table 3. An equation was written using the regression coefficients and the constant (which is not shown in the table). The mean for gunner's AFQT was substituted and the mean for average cuts subtracted to yield an equation expressing average cuts in deviation score terms as a quadratic function of gunner's time in position. This was multiplied by 100 to give an estimated deviation score for total Table VIII cuts. A similar process was used to construct an equation expressing estimated Table VIII total score, again in deviation terms, as a quadratic function of gunner's time in position.

Using these two equations, two curves were plotted in Figure 1 to show how much tank crews with gunners of varying amounts of experience would be expected to deviate above or below the mean performance across crews. The

²Battalion membership is presented as a coded vector of 18 terms, one less than the number of battalions. Each term has its own weight and in sum they describe the battalion differences. These weights are not presented because the specific differences between the battalions in our sample are not expected to generalize beyond the sample. Only the strength of their effects is important.

curvilinear trend, most pronounced for estimated total score, is apparent in both curves. For estimated score, the curve even shows a trend toward decreasing performance at the upper ranges of experience. The maximum predicted score occurs at approximately 22 months' experience. The minimum for estimated cuts occurs at approximately 40 months' time in position. Also apparent from the figure is that gunner's experience has a greater impact on score than on procedural cuts. Although cuts is one component of score, it is apparently not the only component through which experience effects are mediated. The correlations in Table 1 shows that opening time, another component of score, is related to experience. Presumably, experience also affects total score by its effect on opening time. Finally, the figure shows that experience is reliably predictive of approximately a 50 point spread in Table VIII scores and a 10 point spread in cuts. While crews with low experience gunner's are not expected to fail, everything else equal, experience is associated with up to a 50 point difference in score.

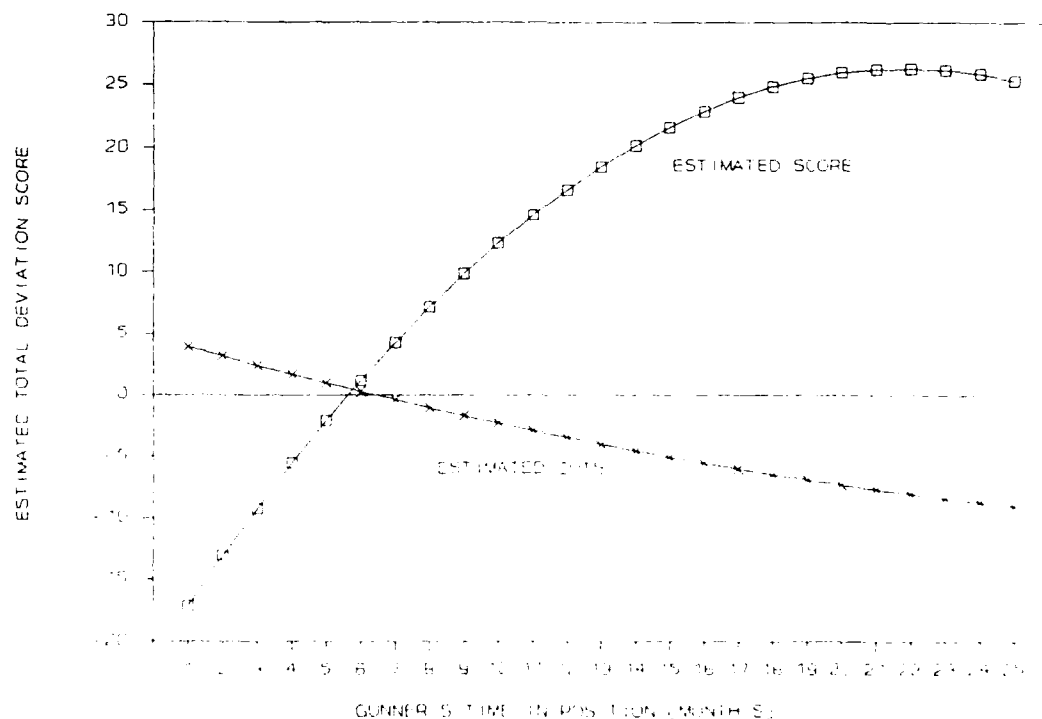


Figure 1. Estimated crew deviations from mean Table VIII score and cuts as functions of gunner's time in position.

Table 4 presents the regression results for hit probability. This analysis began with only battalion membership and gunner's Combat score. Both are retained in the regression solution.

Finally, an analysis was conducted of the potential joint contributions to opening time of battalion membership, tank commander's time in position (linear and quadratic) and gunner's time in position (linear and quadratic).

Tank commander's experience is confounded with gunner's experience (correlation between the two is .31), and has no unique contribution. If we adhere to the traditional .05 significance level, the multiple regression solution reduces to the zero order correlation between gunner's experience and average opening time.

In addition to these additive models, interactions between experience and ability were also examined within and between tank commander and gunner. No interactions were detected.

Table 4

Composite Effects of Battalion and Gunner ASVAB Combat Score on Table VII Total Hit Probability

Source	Sum of Squares	DF	Mean Square	F-Ratio	p	Coefficient	Unique Variance
Battalion	.598	18	.033	2.61	.001	----	11.1%
Gunner Combat Score	.061	1	.061	4.80	.03	.001	1.1%
Error	4.740	372	.013				Total: 12.2%
				R = .354	R ² = .126	p < .01	

Summary. This research has shown that tank commander's experience, gunner's experience and gunner's cognitive ability are related to tank crew performance on Table VIII. Although small, these relationships are rather remarkable given: (1) We are predicting crew performance from individual characteristics. (2) Performance from one part of Table VIII to another is not predictable (e.g. day engagements to night engagements, offensive engagements to defensive engagements; Hoffman, 1988).

References

- Department of the Army. (1986). Tank Combat Tables M1 (Field Manual 17-12-1). Washington, DC: Author.
- Hoffman, R. G. (1988). Grafenwöhr Tank Table VIII: Descriptive statistics (HumRRO Interim Report). Ft. Knox, KY: Human Resources Research Organization.
- Hoffman, R. G., & Morrison, J. E. (1988). Requirements for a device-based training and testing program for M1 gunnery: Volume 1. Rationale and Summary of Results (ARI Technical Report 783). Alexandria, VA: U. S. Army Research Institute for the Behavioral and Social Sciences.
- McDaniel, M. A., Schmidt, F. L., & Hunter, J. E. (1988). Job experience correlates of job performance. Journal of Applied Psychology, 73, 327-330.

THE UNIFIED TRI-SERVICES COGNITIVE PERFORMANCE ASSESSMENT BATTERY

G. Rufus Sessions, David R. Thorne,
Samuel L. Moise, Jr., and Frederick W. Hegge

Walter Reed Army Institute of Research¹
Washington, D.C. 20307
and Ergometrics Technology, Inc.
Los Gatos, CA 95030

Introduction

The Unified Tri-Service Cognitive Performance Assessment Battery (UTC-PAB) is the primary instrument for the assessment of cognitive performance in a multiple level drug evaluation program established by the Joint Working Group on Drug Dependent Degradation in Military Performance (JWGD³ MILPERF), a joint services research and development program sponsored by the U.S. Army Medical Research and Development Command and under the execution authority of the Walter Reed Army Institute of Research. The UTC-PAB consists of a computerized test authoring and test administration system, 25 individual tests, and supporting documentation. This paper reviews the history of the UTC-PAB construction, rationale and criteria for test selection and the structural framework within which the tests are organized. Test authoring and execution within the computerized UTC-PAB is centered around the concept of the Generic Task Interval. This concept is discussed in relation to its utility in test construction for UTC-PAB or for new tasks or tests.

Joint Working Group on Drug Dependent Degradation in Military Performance

The work carried out in the JWGD³ MILPERF program and reflected in the development of the UTC-PAB is an extension of safety screening. Drugs entering the program have received prior efficacy and direct toxicological screening. They are safe for human use in accordance with standard drug development practice. The goal of this program is to extend safety assessment to risks that may arise not from the drug itself but from the interactions between drug actions, mission requirements, and mission setting.

Multicenter safety screening requires procedural standardization such as: a. drug dosage and regimen; b. content and conduct of investigational procedures; and c. data analysis and reporting. The JWGD³ MILPERF program has developed screening techniques and

¹This material has been reviewed by the Walter Reed Army Institute of Research, and there is no objection to its presentation and/or publication. The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the Department of the Army or the Department of Defense.

instruments designed to achieve such standardization in extended drug safety screening studies.

Drugs designated for testing may enter the program with a long history of medical use for other purposes, or they may be new drugs that do not have an extensive history of use. In either case the drugs will be subjected to an initial phase of testing referred to as Basal, or Level I Testing. This screening phase is essentially that of the controlled clinical investigation. The program elements used in this phase resemble clinical diagnostic instruments and form the basis of the initial assessment of drug effects. The tests used in this phase require little pretraining and short times to administer. The tests to be utilized in the screen include neuropsychological, neurophysiological and psychomotor test batteries. The JWGD³ MILPERF Neurophysiological Test Battery is described elsewhere (Reeves et al., in preparation). The Basal Testing Screen is intended to fill gaps in our knowledge of drug effects and to provide the experimental base for further testing.

Information from the data base compiled from the literature and Basal Testing Screen is used to select a candidate set of performance assessment instruments and tests tailored to the drug being screened. This represents the first research grade performance assessment. The elements of the Performance Testing, or Level II Screen include the JWGD³ MILPERF Physical Performance Assessment Battery (Montgomery et al, in preparation) to assess drug effects on physical parameters including strength, endurance and dexterity, and the UTC-PAB to assess drug effects on cognitive performance. The final phase of the screening program involves performance testing in simulated or synthetic work environments in conjunction with stressors such as heat, cold, fatigue, etc. (Level III testing).

History of UTC-PAB Development

For a full treatment of the history of UTC-PAB development, the rationale and criteria for test selection and the organizing framework for the individual tests, see Englund et al. (1987). For information on test purposes, literature review, technical descriptions, validity, sensitivity, reliability, data specifications and training requirements see Perez et al. (1987).

Test selection for the UTC-PAB occurred during a three-day workshop sponsored by the JWGD³ MILPERF in November 1984. Researchers from participating Army, Navy and Air Force laboratories attended. Each attendee had a background in performance or workload assessment and information processing and most were intimately involved in the development of service specific computerized batteries. The workshop objectives were: a. to select tests from existing inventories that met guidelines for inclusion in the battery, and b. to agree upon the formats for each test. As a result of this workshop a twenty-five item menu of tests was designated as the UTC-PAB. The product of this workshop provided the basis for the subsequent development of a software plan which capitalized on the structural similarities between tests.

Individual performance tests were chosen that would contribute

to a common menu of available tests from which investigators could select and tailor a subset of tests for particular applications. Tests selected were required to be backed by prior research, including measures of construct validity, reliability, and sensitivity. No attempt was made to select tests on the basis on any single theoretical model; selection decisions were based primarily on the potential utility of individual tests in drug screening.

Standardization of cognitive performance assessment tests, their administration and analysis was a fundamental rubric for the UTC-PAB development. The JWGD³ MILPERF program requires standardization to ensure compatibility between interlaboratory research efforts in coordinated drug screening studies, and to provide the basis for direct comparison and pooling of data collected from the different laboratories involved in the program. Flexibility and room for growth was also designed into the UTC-PAB. While the menu of twenty-five tests provides a "core" set of tests for selection into specific batteries, provisions were made for inclusion of additional tests or creation of new tests within the same test environment.

Structural Framework for Classification of UTC-PAB Tests

The structural framework for categorizing tests in the UTC-PAB has been detailed by Englund et al. (1987). This framework was intended to provide guidelines for selecting subsets of tests from the battery for specific research applications:

At a general level, tasks in the battery can be classified according to the type of information processing function most heavily involved in their performance. Although a variety of such functions are involved in task performance, two dimensions of processing that are particularly critical to assessment of drug effects are: (a) the stage of information processing which is most markedly affected by the demands of the task, and (b) the requirement to divide or selectively employ attentional capacity among sources of information. Several major functions can be distinguished within the stages of processing dimensions. These include perceptual input function, such as information detection and identification; central processing functions, including a variety of memory and information integration/manipulation functions; and motor output or response execution functions. Integration and manipulation functions within central processing can be further subdivided into those based on linguistic/symbolic forms of information versus those involving spatial information. The dimensions of processing within this framework are consistent with several current theoretical descriptions of the human information processing system, including those of Wickens (1984) and Shingledecker (1984). (Englund et al., 1987, p. 4)

The UTC-PAB Tests

The following is the list of UTC-PAB test elements, categorized according to the framework outlined above:

I. STAGES OF PROCESSING

A. Perceptual Input, Detection, and Identification

Visual Scanning
Simultaneous Pattern Comparison
Visual Probability Monitoring
Four-Choice Reaction Time
Alphanumeric Vigilance

B. Central Processing

1. Short & Long Term Memory, Associative Memory

Auditory Sternberg
Visual Sternberg
Continuous Recall
Item Order
Code Substitution

2. Information Integration/Manipulation (Linguistic/Symbolic Mode)

Linguistic Processing
Mathematical Processing
Two Column Addition
Grammatical Reasoning (Traditional)
Grammatical Reasoning (Symbolic)

3. Information Integration/Manipulation (Spatial Mode)

Spatial Processing
Matrix Rotation
Matching to Sample
Manikin Test
Successive Pattern Comparison
Time Wall

C. Output/Response Execution

Interval Production
Critical Instability Tracking

II. SELECTIVE/DIVIDED ATTENTION

Dichotic Listening
Stroop
Sternberg Memory-Tracking Combination

The Computerized UTC-PAB

The UTC-PAB software is comprised of two separate programs, a Configuration System for authoring tests and batteries, and a Runtime System for test execution and data collection. The Configuration System is an interactive, integrated, menu-driven programming system that permits the creation of test batteries, with individual tests and test menus, without the need for programming skills. The Runtime System is similarly an interactive, integrated, menu-driven program that permits the selection and administration of test batteries by the researcher or by paraprofessionals.

The authoring system is designed around a uniform user interface permitting the full definition of a subject task, the combination of tasks into tests, and the aggregation of tests into batteries. Provision is made for training modules with context sensitive help. An EGA graphics editor and two channels of synthesized voice provide for the creation of visual and auditory stimuli.

The UTC-PAB software is written in the C programming language and is designed to operate on IBM AT compatible microcomputers equipped with 640KB RAM memory operating under MS-DOS 2.0 or above. High-resolution RGB or EGA color monitors provide video display. Additional hardware requirements include a 80286 math co-processor and a separate laboratory interface board to provide timing functions and analog and digital interfacing to an optional subject workstation containing multiple response panels. The computer keyboard keys can be used for subject response input as well. Complete technical descriptions of Hardware/Software design and specifications for the UTC-PAB can be found elsewhere (Reeves, et al., in preparation).

The Generic Task Interval

The UTC-PAB test authoring software is based on an interval model of performance assessment. Nearly any task that can be described in terms of a sequence of intervals can be implemented with the system.

The Battery is the highest level of the system. It consists of a list of Tests, their presentation sequence, and inter-test intervals. Tests are a collection of tasks, their presentation sequence and inter-task intervals. Examples of tests might be the Manikin Test or the Stroop Test. Such tests are composed of repeated sequences of tasks. The task itself is the heart of a given test and its structure defines the stature of the test. A task might represent a given trial of a particular test and may be composed of a single or a series of stimulus-response sequences, or intervals. The Interval is the lowest level of the system, and is composed of visual or auditory stimuli, the delays between presentations of these stimuli, the range of allowable (appropriate) responses, the specification of correct and incorrect responses, and the sequence of responses allowed in the interval. An interval might consist of the presentation of textual information on the screen for a specified time period or until the subject makes a specified response.

The program is structured to incorporate several different types of intervals representing the distinctly different types of stimulus-response sequences that a tester might want to use in a task. For example, a common task that a tester may want to present to a subject might involve the following sequences: (1) Requiring the subject to fix his attention to a screen or listen for a sound; (2) Presenting the subject with information (visual or auditory stimuli); (3) Requiring the subject to remember the information for some period of time; (4) Presenting the subject with another stimulus and asking him to give information or make a response based on some comparison or rule; (5) Presenting the subject with feedback concerning the correctness of the response.

The user interface of the Configuration System is designed to aid the experimenter in carefully defining the characteristics and parameters of each component of a task in a test. The system prompts the user through the necessary steps to design the stimuli involved in the intervals, to define the timing relationships between the stimuli and the various possible responses of the subject, and to provide a means of determining the correctness of the response, and, if necessary, to provide feedback to the subject based on the consequences of his answers or responses.

The UTC-PAB offers researchers enormous power and flexibility for constructing and executing arbitrarily complex applications in automated cognitive performance assessment studies without the requirement of extensive programming.

References

- Montgomery, L. C., Kyle, S. B., Smoak, B. L. & Deuster, P. A. Performance physiology test battery for chemical warfare treatment/Pretreatment drugs (JWGD³ MILPERF Report, in preparation).
- Englund, C. E., Reeves, C. A., Shingledecker, D. R., Thorne, D. R., Wilson, K. P. & Hegge, F. W. (1987) Unified tri-service cognitive performance assessment battery (UTC-PAB): I. Design and specification of the battery (NHRC Report No. 87-10). San Diego, CA: Naval Medical Research Center.
- Perez, W. A., Masline, P. J., Ramsey, E. G. & Urban, K. E. (1987). Univied tri-services cognitive performance assessment battery: Review and methodology (AAMRL Report No. TR-87-007). Dayton, OH: Armstrong Aerospace Medical Research Laboratory. (DTIC No. AD-A181-697).
- Reeves, D. L., Winter, S. L., Thorne, D. R. & Hegge, F. W. Unified tri-service cognitive performance assessment battery (UTC-PAB): II. Hardware/Software Design and Specifications (JWGD³ MILPERF Report No. 89-1, in preparation).

RE-EVALUATING PRACTICAL PERFORMANCE

ASSESSMENT STRATEGIES IN THE CANADIAN NAVY

by

LCdr Edward G. Barnett, CF

Canadian Forces Fleet School Halifax

Introduction

The Canadian Navy is currently experiencing rapid technological growth as it procures more modern and sophisticated equipment. The traditional naval training system has been hard-pressed to keep pace with these advances. Increasing emphasis has been placed on developing generic strategies to implement a more realistic, flexible systems approach to training. This has been particularly true with respect to performance testing procedures as practised in the Combat Systems Engineering Division (CSED), Canadian Forces Fleet School Halifax (CFFSH).

This paper describes a performance testing methodology that was designed for several naval technical occupations within CSED. Its intended purpose was to provide a more effective and efficient, standardized performance testing process for both generic and equipment specific procedures. The development of this strategy is described as are its principle tenets and intended applications.

Background

Baker and O'Neil (1987, p. 343) clearly point out that the introduction of new and exciting educational/training technologies has often overshadowed the serious requirement for matching assessment strategies. This was certainly the case in CSED. Performance tests of the simulated performance and work sample types were generally of very poor quality. The majority were unreliable, unusable and often invalid.

A special project team was organized in late 1985 to analyze and develop practical solutions to remedy this situation. It became readily apparent that a revised performance test development process was required.

Test Development Procedure

Though many renowned test development authorities, such as Chase (1978), Gronlund (1982) and Popham (1981) have all described methodologies to develop solid performance tests, few have provided the level of detail required by the untrained test developer to be effective. The majority of CSED naval technicians employed in training development activities fell into this category.

A practical set of steps were developed to remedy this situation and are briefly outlined as follows:

1 Phone Contact: (Centrex) 1-902-427-8178

- a. Step 1 - analyze the performance objective;
- b. Step 2 - prepare a list of test steps/items for product and or process assessment;
- c. Step 3 - produce a test prescription qualifying the specific behaviour that is required of the trainee for each step/item;
- d. Step 4 - translate the test prescription onto a test sheet with all specific behaviours qualified; and
- e. Step 5 - insert the completed test sheet into the standardized test format (intro page, admin instructions, pass/fail criteria etc...).

Figure 1 represents an example of the completed test prescription process (Step 3) for one major step in a diagnostic procedure. Each sub-step produced from the job oriented, performance objective analysis is reviewed. A determination is made as to the appropriate trainee behaviour that will confirm to the tester that the sub-step has been successfully achieved. This will generally be either a practical/skill related behaviour or a verbal one, though other possibilities exist.

Figure 2 illustrates the completed test prescription translation process (Step 4), using the Figure 1 example. Specific behaviours are now qualified for each sub-step with any necessary comments included. The test sheet is now complete and ready for use as part of the overall test package. Though this is a relatively simple illustration, it represents a more precise, standardized test development procedure than ever employed in CSED before.

Advantages

The advantages of such a test development philosophy are numerous. Increased content validity is evident as specific performance indicators are more precisely defined and reflected in the test sheets. Inexperienced test developers have a more usable strategy to employ and test administrators have an easier format to follow. Increased reliability is provided via a more consistent test development and testing structure. Lastly, this type of test production lends itself well to an automated, word processing format allowing for more rapid test production and revision.

Conclusion

Preliminary implementation of this methodology has been very well received by all CSED personnel currently employed in test development initiatives. Test developers have been extremely pleased with the simplicity yet effectiveness of the design. Future research will involve analyzing test production times to see if there has been an increase in efficiency and effectiveness using this strategy.

FIGURE 1

CHECK/TEST PRESCRIPTION

Performance Statement: Perform Diagnostics On The
AN/ULQ - 6C Countermeasure Set

Equipment/Unit(s) Being Checked or Tested On:

- a. AN/ULQ - 6
 b. _____
 c. _____
 d. _____
 e. _____

<u>List of Steps or Items</u>	<u>Na</u>	<u>Prac. Verb.</u>			<u>Comments</u>
		<u>Perf.</u>	<u>Perf.</u>	<u>Other</u>	
1. <u>Identified Overall</u>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<u>Major Step</u>
<u>Nature and Location of Fault</u>					
a. <u>observed correct</u>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
<u>operative sequence</u>					
b. <u>identified location</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
<u>of fault</u>					
c. <u>identified type</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	
<u>of fault</u>					
d. <u>observed all</u>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<u>Critical Factor</u>
<u>safety precautions</u>					
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

FIGURE 2

C. CHECK/TEST REPORT FOR STEPS OR ITEMS

- ☒ Process Check/Test (procedures or steps in the process)
☐ Product Check/Test (characteristics of finished product)
☐ Both

Performance Statement(s): Perform Diagnostics On The
AN/ULQ - 6 C Countermeasure Set

Equipment/Unit(s) Being Checked or Tested On:

- a. AN/ULQ - 6
b. _____
c. _____
d. _____
e. _____

<u>List of Steps or Items</u>	<u>Yes</u>	<u>No</u>	<u>Other</u>	<u>Comments</u>
1. <u>Identified Overall Nature</u> <u>and Location of Fault</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<u>Major Step</u>
a. <u>observed correct</u> <u>operative sequence</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
b. <u>verbally identified</u> <u>location of fault</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
c. <u>verbally identified</u> <u>type of fault</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	
d. <u>observed all safety</u> <u>precautions</u>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<u>Critical Factor</u>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	

References

Baker, E.L. & O'Neil, H.F. (1987). Assessing instructional outcomes. In R.M. Gagné, (Ed.). Instructional Technology: Foundations (1st Ed.)(pp.343-377). Hillsdale, NJ: Lawrence Erlbaum Associates.

Chase, C.I. (1978). Measurement for Educational Evaluation (2nd Ed.). Reading, MA: Addison-Wesley Publishing.

Gronlund, N.E. (1982). Constructing Achievement Tests (3rd Ed.). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Popham, W.J. (1981). Modern Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall, Inc.

A FULLY AUTOMATED MEMORY AND SEARCH TASK

Charles A. Salter

Laurie S. Lester

Heather Dragsbaek

Richard D. Popper

Edward Hirsch

U. S. Army Natick Research, Development & Engineering Center

ABSTRACT

We developed a fully automated Memory and Search Task (MAST) for use on small hand-held computers weighing less than a pound each. This replaces paper and pencil versions of the MAST that, although useful in documenting performance changes due to fatigue and circadian rhythm, were difficult to use and score under field conditions. A randomly generated stimulus sequence of 16 alphabetic characters is presented on the screen along with a target sequence of two, four, or six characters. For two minutes under each target-length condition the subject responds by pressing a key indicating whether he thinks the target can or cannot be found in the stimulus. This task was tested in a classroom with three groups of soldiers at Fort Devens, MA. One group practiced the task twice a day at 1300 hours and showed a typical learning curve over five days. A second group performed the task ten times in a row, starting at 1300 hours, and revealed that massed practice is, as expected, less efficient at improving performance on this task. A third group performed the task three times a day, at 1000, 1300, and 1430 hours, twice each time, for five days. Once initial learning was complete, this group tended towards a significant post-lunch slump in performance, followed by significant improvement later in the afternoon ($P < .05$). These results indicate that the new automated MAST could prove to be a sensitive measure of performance in the laboratory, and because of its portability may prove useful in the field as well.

INTRODUCTION

There are a plethora of human performance tasks which can be administered in a laboratory or garrison setting using paper-and-pencil measures, special test equipment, or computers. However, many of these systems are not portable enough for use in a military setting in the field, with troops constantly on the move in a simulated combat scenario where experimenters can not always be present (Popper et al, 1988). In practice, this means that field studies often use performance measures that are different from those preferred in the laboratory. The use of different measures makes many of the two kinds of studies fundamentally non-comparable. Thus findings in the field may not apply to the laboratory situation and vice versa. To help overcome this problem, we developed a small, portable, sturdy, yet fully automated performance system which can be used equally well under laboratory, garrison, and field conditions.

THE COMPUTER SYSTEM

Our task was programmed to run on a Sharp 1500A computer (see Figure 1). Table 1 lists the specifications for this machine. To protect the device under field conditions, a special case was fabricated for it. The case includes an inch of foam padding which completely surrounds the computer when not in use and a strap by which it can be fastened to a belt or pack.

THE PERFORMANCE TASK

Using the BASIC computer language, we programmed the Sharp 1500A to run the Memory and Search Task (MAST). This task was originally developed as a paper-and-pencil test by Kaplan et al (1966) to measure visual search and immediate memory skills. The MAST was modified by Folkard (1976) so that short term memory loading could be varied, but remained a paper-and-pencil task. We were the first to produce a fully automated version of it for handheld computers like the Sharp. The task involves looking at a randomly generated stimulus sequence of 16 letters which are not in any meaningful order, then looking at a target of either 2, 4, or 6 letters and deciding whether that target can be found in any order in the original stimulus. For example, one trial might consist of "GEUUKQJMYYPQWPOH EJ". The subject then simply presses "Y" for yes or "N" for no. In the example given, both E and J from the target can be found in the stimulus, so the correct response would be "Y". If only part of the target, but not all of it, can be found in the stimulus, then the correct response is "N."

To complete one iteration of the task, the subject enters the last four digits of his social security number and runs through two minutes with a two-character target, two minutes with a four-character target, and two minutes with a six-character target, though the sequence of these three conditions is varied randomly each time. For the entire two minutes the target remains the same, but the sixteen-character stimulus changes randomly after each response. The subject is instructed to work as quickly and accurately as he can. At the end of one six-minute iteration, the computer screen provides feedback for each condition separately on the number of trials completed, the correct hits, the misses, the false alarms, and the correct rejections. After two different iterations, approximately 75 minutes is required for the program to generate new random sequences for the next pair of iterations. The computer will automatically record and retain the data of up to 60 complete iterations of the task along with the date and time each was initiated. This feature makes the system ideal for long-term field studies in which a number of iterations under different conditions are desired. At completion of the study, the data can be uploaded directly to a larger computer for data analysis. Alternatively, the data can be viewed sequentially on the Sharp 1500A screen itself or printed on paper using a Sharp CE-150 hand-held printer.

TABLE 1
SHARP 1500A SPECIFICATIONS

ITEM	SPECIFICATIONS
SIZE	195mm WIDE X 25.5mm HiGH X 86mm DEEP
WEIGHT	375g
CPU	C-MOS 8-BIT CPU
ROM	16K
RAM	8.5K (STANDARD); 24.5K (OPTIONAL)
KEYBOARD	65 KEYS (ALPHABET, NUMERIC, FUNCTION, AND SOFTWARE KEYS)
DISPLAY	7 X 156 MINI-GRAPHIC DISPLAY; 26-DIGIT LIQUID CRYSTAL DISPLAY
POWER	4 AA NON RECHARGEABLE BATTERIES

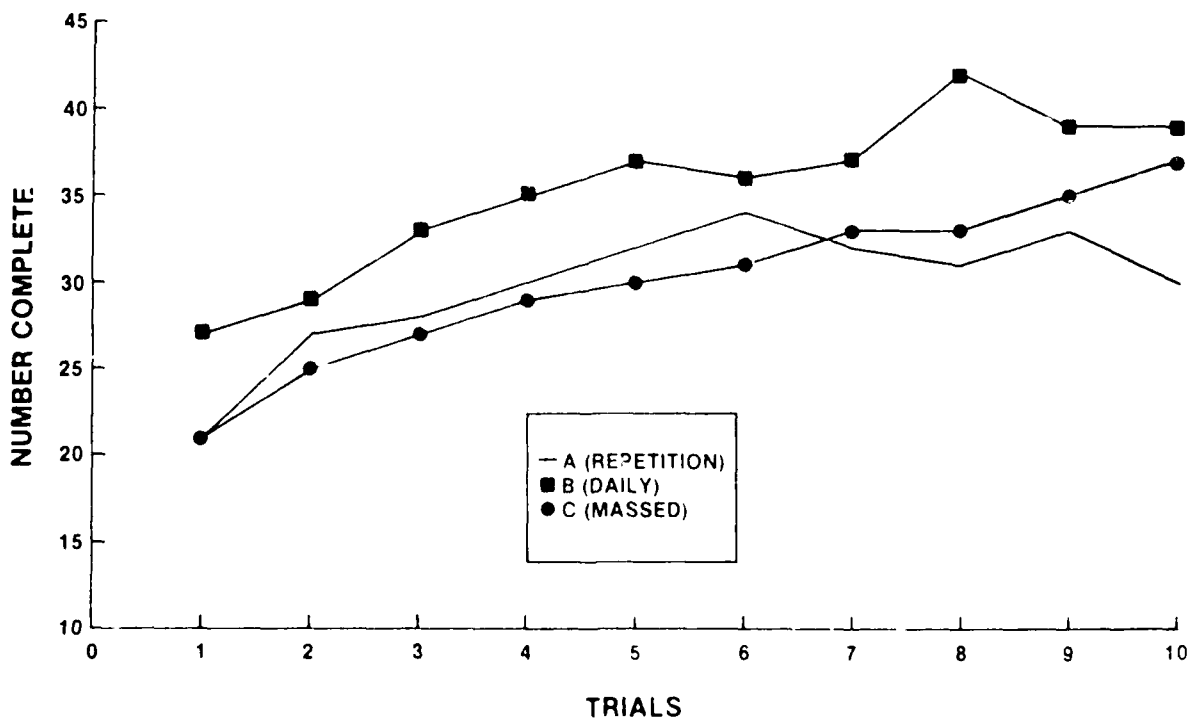
FIGURE 1
THE SHARP 1500A IN AN OPENED MILITARY CASE



STUDIES INVOLVING THE MAST

The initial study of the performance characteristics of this automated task was conducted at Ft. Devens, MA with three groups of soldiers from the 112th Military Intelligence Brigade. These soldiers were tested in a classroom setting on post. Each group started with 8 or 9 soldiers each, but personnel who were unable to complete all the required days of testing were dropped from analysis. Group A (6 subjects) met at 1000, 1300, and 1430 hours each day for five days, performing two iterations of the MAST each time, for a total of 30 iterations each. Group B (6 subjects) met at 1300 each day for five days, performing the MAST twice each time, for a total of 10 iterations each. Group C (7 subjects) met at 1300 only one day, and each completed a set of 10 iterations all at once. Subjects were instructed to eat a normal breakfast and lunch, but to eat lunch only between 1130 and 1200 hours on testing days.

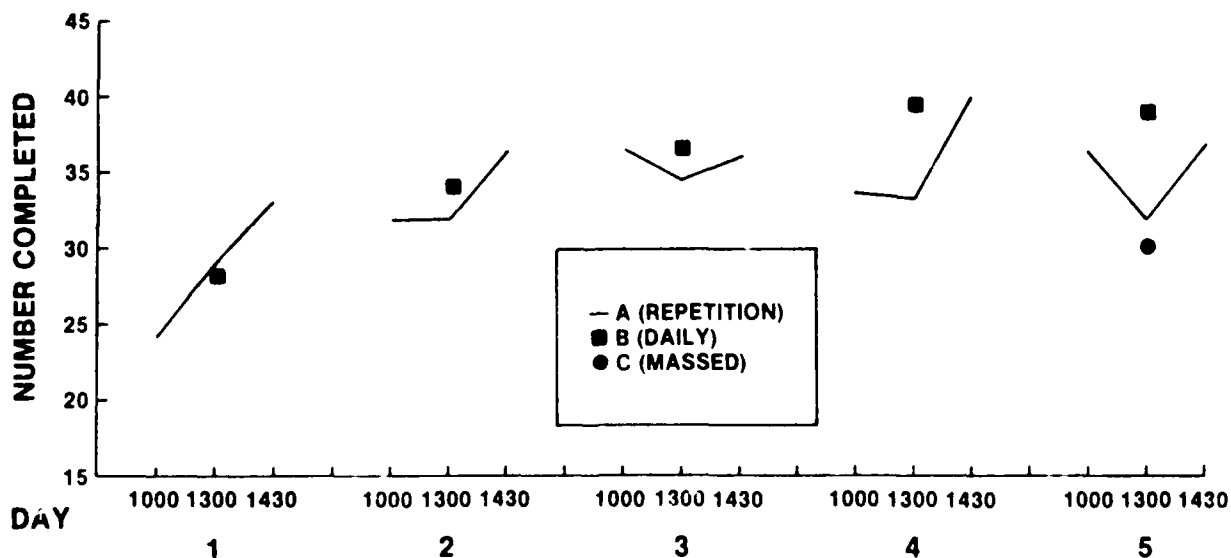
FIGURE 2
LEARNING CURVE FOR FIRST TEN ITERATIONS
TWO CHARACTER TARGET CONDITION



Some illustrative results on the first ten iterations with the two-character target condition can be seen in Figure 2. Group B (the daily group), which performed the fewest trials per day, appeared to do the best. Group C (the massed practice group), which performed the most at once, appeared to do worse, though even there a trend towards improvement with practice can be seen. These group differences, though nonsignificant statistically, are in the same direction as the established finding that distributed practice is superior to massed practice (Swenson, 1980).

Group A (the repetition group) was the only group to perform the task at multiple times per day. We timed their repetition schedule so that we could check for possible effects such as the "post-lunch slump" (Craig, 1986). Figure 3 shows the results for the two-character target condition by time of day. (Each data point represents the means of the two iterations at that time). On the first day, Group A appeared to show the typical

FIGURE 3
PERFORMANCE BY TIME OF DAY
TWO CHARACTER TARGET CONDITION



upward learning curve. But on subsequent days, that group appeared to do poorest an hour after lunch, doing better both before lunch and later that afternoon. These differences between the post-lunch and later-afternoon times were statistically significant ($P < .05$), suggesting that performance on the MAST does show a post-lunch slump followed by a later recovery.

ONGOING STUDIES

We are currently running other studies with the MAST. For instance, we are conducting a study of simulated jet lag in time-isolated apartments. Subjects spend six days at normal time establishing their baseline on the MAST and other performance and physiological measures. On the sixth night, subjects are awakened without warning six hours before their normal time, thus simulating an easterly time shift of six hours. For the next nine days, they are kept to the new "time zone," and we are monitoring performance disruptions caused by this shift and subsequent recovery afterwards, using the MAST as well as other measures.

Depending on the results of this and other laboratory studies, we plan to take the MAST on field studies of actual jet lag, testing the performance of US Marines before, during, and after trans-Atlantic flights. The portability of this system should make it ideal for such field studies.

References

- Craig, A. (1986). Acute effects of meals on perceptual and cognitive efficiency. Nutrition Reviews/Supplement, 44, 163-171.
- Folkard, S., Knauth, P., and Monk, T. H. (1976). The effect of memory load on the circadian variation in performance efficiency under a rapidly rotating shift system. Ergonomics, 19, 479-488.
- Kaplan, I. T., Carvellas, T., and Metlay, W. (1966). Visual search and immediate memory. Journal of Experimental Psychology, 71, 488-493.
- Popper, R., Dragsbaek, H., Siegel, S. F., and Hirsch, E. (1988). Use of pocket computers for self-administration of cognitive tests in the field. Behavior Research Methods, Instruments, & Computers, 20, 481-484.
- Swenson, L. C. (1980). Theories of Learning. Belmont, CA: Wadsworth.

Armed Services Vocational Aptitude Battery and Vehicle Identification Performance Relationships

Otto H. Heuckeroth

Norman D. Smith

ARMY RESEARCH INSTITUTE, FORT HOOD FIELD UNIT

Current Army-wide and executive emphasis on MANPRINT to aid in the planning, evaluation and development of Army systems stresses the importance of understanding the relationship between soldier performance, aptitude and training achievement. In this spirit the purposes of the present paper are to: 1) Explore the relationship between one criterion performance measure--vehicle identification accuracy--with the Armed Services Vocational Aptitude Battery (ASVAB) and Subtest scores as predictor variables; and 2) to present additional analyses which document the validity of those relationships. For brevity in the presentation today, only the results involving ASVAB Scaled Scores are discussed.

From 1980 to 1986 the Fort Hood Field Unit conducted approximately 13 independent research efforts in the Target Acquisition Analysis and Training System (TAATS) program. Each of those efforts focused on design, development and testing of a series of Combat Vehicle Identification (CVI) programs. In eleven of those efforts, soldiers made vehicle identification responses to a common set of photopic (daylight) vehicle images independently presented. The number of images correctly identified was the criterion (dependent) variable. In each case, images of the same vehicles in the same views (front and oblique) were used. To address these purposes, four analytic techniques were used: 1) Correlations of individual ASVAB Composites and Subtests with soldier vehicle identification performance, 2) multiple correlations between soldier vehicle identification performance and ASVAB Composites or Subtests as predictor variables, 3) correlations of differentially weighted ASVAB predictor variables with vehicle identification performance, and 4) discriminant analyses.

It is important to note that in this effort separate analyses were generally performed using data from ASVAB Test Forms 5-7, 8-14 and 5-14 for random halves of soldiers for which data were available, as well as for the entire set of data available. Separate analyses for different ASVAB Test Forms were motivated by the general understanding that there are rather major differences in test structure for ASVAB versions beginning with Test Form 8 compared with earlier ASVAB versions. Analyses for random halves were completed primarily to address the validity of findings reported. The reliability of the criterion measure was .88.

RESULTS

Individual ASVAB Scaled Scores Correlations with Identification Performance

Table 1 presents the correlations between each ASVAB Scaled Score and vehicle identification performance together with their statistical significance, sample size (N) and for ODD/EVEN halves a Z statistic to assess the significance of differences between correlations. Including the correlations for ODD and EVEN halves as well as those based on the total sample, all but three are significantly different from zero; all of the

Table 1

Correlational Matrix of Identification Performance With ASVAB Scaled Scores for Independent Sample Halves and Total Sample for Soldiers Who Took an ASVAB Test Form 5-7 or 8-14

ASVAB SCALED SCORES	TEST FORM 5-7 ODD HALF	TEST FORM 5-7 EVEN HALF	TEST FORM 5-7 TOTAL	TEST FORM 8-14 ODD HALF	TEST FORM 8-14 EVEN HALF	TEST FORM 8-14 TOTAL	TEST FORM 5-14 ODD HALF	TEST FORM 5-14 EVEN HALF	TEST FORM 5-14 TOTAL
AFQT P N Z ^a	.290 .0002 155	.370 .0001 154	.330 .0001 309	.217 .0112 136	.297 .0004 136	.254 .0001 272	.325 .0001 291	.289 .0001 290	.307 .0001 581
		.70			.69			.67	
CO P N Z	.290 .0003 155	.369 .0001 155	.330 .0001 310	.249 .0033 137	.218 .0108 136	.233 .0001 273	.292 .0001 292	.311 .0001 291	.302 .0001 583
		.78			.27			.23	
FA P N Z	.290 .0002 156	.363 .0001 155	.323 .0001 311	.171 .0461 136	.290 .0006 136	.227 .0002 272	.278 .0001 292	.311 .0001 291	.296 .0001 583
		.71			1.03			.44	
MM P N Z	.310 .0001 157	.323 .0001 156	.315 .0001 313	.160 .0808 138	.282 .0008 138	.217 .0003 276	.318 .0001 295	.256 .0001 294	.287 .0001 589
		.12			1.06			.81	
GM P N Z	.329 .0001 156	.390 .0001 155	.357 .0001 311	.395 .0001 138	.174 .0425 137	.282 .0001 273	.309 .0001 293	.363 .0001 293	.336 .0001 586
		.61			1.99*			.73	
CL P N Z	.252 .0016 155	.238 .0028 155	.246 .0001 310	.045 .6002 137	.205 .0162 137	.119 .0484 274	.221 .0001 292	.200 .0006 292	.212 .0001 584
		.14			1.33			.26	
GT P N Z	.229 .0044 154	.351 .0001 153	.285 .0001 307	.304 .0001 156	.182 .0234 135	.235 .0001 311	.283 .0001 309	.241 .0001 309	.261 .0001 618
		1.16			1.14			.36	
EL P N Z	.288 .0003 156	.406 .0001 155	.342 .0001 311	.278 .0011 136	.286 .0007 136	.278 .0001 272	.325 .0001 292	.327 .0001 291	.326 .0001 583
		1.17			.07			.02	
SC P N Z	.375 .0001 155	.334 .0001 155	.354 .0001 310	.210 .0136 138	.193 .0236 137	.202 .0008 275	.336 .0001 293	.272 .0001 292	.305 .0001 585
		.42			.16			.84	
ST P N Z	.300 .0001 156	.336 .0001 155	.317 .0001 311	.419 .0001 138	.173 .0438 137	.289 .0001 275	.282 .0001 293	.349 .0001 293	.316 .0001 586
		.35			2.21*			.90	
OF P N Z	.228 .0042 157	.366 .0001 156	.290 .0001 313	.163 .0535 138	.308 .0002 138	.235 .0001 276	.292 .0001 295	.272 .0001 294	.282 .0001 589
		1.32			1.26			.27	

^aP values address the significance of individual correlations and were provided as part of the Statistical Analyses Software (SAS) PROC CORR output.

^bZ values were computed by the formula: $|z_{r_1} - z_{r_2}| / \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$. See McNemar, Q. *Psychological Statistics*.

John Wiley and Sons, Inc. 1962, pp. 139-140. For two tailed tests, $|Z| \geq 1.96$ is significant at $p < .05$. Tabled Z values address the significance of difference between correlations obtained using independent halves of available data. Z values which are significant are noted by *.

correlations based on the total sample are statistically significant. The range of correlations obtained using test forms 5-7 is .246 - .354; for test forms 8-14 the range is .119 -.289. For all test forms the range is .212 - .336. This means that between 1.4% and 12.5% of the variability between ASVAB Scaled Scores and Identification performance is in common.

In addressing the validity of findings presented in Table 1 correlations of random halves are included in this table. For the thirty-three Z test comparisons, only two (approximately 6%) are significant at the .05 level or better. Since there were no a priori reasons to believe any of these differences would be significant we have concluded that the reported relationships are probably valid.

Multiple Correlations Involving ASVAB Scaled Scores and Identification

Performance

Since with eleven predictor variables there are 2047 possible predictor sets, PROC STEPWISE from the SAS package was used to select the best set--one for test form 5-7 data, a second for test form 8-14 data and a third with all test form data. The predictor set selected was the one which satisfied Mallow's criterion--Cp statistic--as described in the SAS manual. These results are presented in row 1 of Table 2. The remaining rows show the multiple correlations for selected combinations of predictor sets. Based on review of Table 2 it appears that the multiple correlations approach does not produce any marked improvement over the individual correlations in the demonstrated relationship between ASVAB Scaled Scores and the criterion variable.

Differential Weighting Correlations

In attempting to determine the relationship among predictive and criterion variables, it seemed reasonable to ask whether weighting different categories of the predictor variables by values other than one might improve the relationship. Inspection of weighted and unweighted correlations in Table 3 indicate that in almost every case using estimated weights for selected categories of ASVAB predictors does lead to increases in the absolute value of obtained correlations; however, only for the OF Composite did differential weighting lead to a statistically larger ($p < .05$) correlational value.

Discriminant Analyses

Early research with the CVI Training Systems indicated large individual soldier performance variability. Further analyses seemed to indicate that soldiers who performed relatively poorly (the lower third) after the first training session (low achievers) showed smaller performance increases with subsequent training. Our research asked whether these "low" achievers could be differentiated from "high" achievers with ASVAB Scaled Scores. High and low achievers were divided into random halves. Odd halves served as the "calibration" sample for a series of discriminant analyses. Table 4 indicates that: 1) High vs low achievers who took an ASVAB form 5-7 could be distinguished about 75% of the time; 2) those who took an ASVAB form 8-14 could be distinguished about 77% of the time; 3) in no case was there a significant difference in classification accuracy between calibration and test samples.

Table 2
Selected Multiple Correlations of Identification Performance With ASVAB Scaled Scores for Independent Sample Halves and Total Sample for Soldiers Who Took An ASVAB Test Form 5-7 or 8-14

Predictor Sets	Test Forms 5-7		Test Forms 8-14		Test Forms 5-14		Total
	Odd Half	Even Half	Odd Half	Even Half	Odd Half	Even Half	
	(SC)	(SC)	(SC From 11 Composites) ^a	(ST)	(OH)	(OH)	(GM From 11 Composites) ^a
R	.431	.286	.361	.279	.293	.413	.361
N	118	118	236	123	242	241	483
Z ^b	1.27			.27	1.50		
	(EL, GT, CL, OF, CO, NM)	(EL, GT, CL, OF, CO, NM)	(EL, GT, CL, OF, CO, NM) ^c	(EL, GT, CL, OF, CO, NM)	(EL, GT, CL, OF, CO, NM)	(EL, GT, CL, OF, CO, NM)	(EL, GT, CL, OF, CO, NM)
R	.466	.365	.400	.323	.343	.402	.360
N	118	118	236	123	242	241	483
Z	.92			.08	.75		
	(ST, CL, SC, CO, OF, AFQT)	(ST, CL, SC, CO, OF, AFQT)	(ST, CL, SC, CO, OF, AFQT) ^c	(ST, CL, SC, CO, OF, AFQT)	(ST, CL, SC, CO, OF, AFQT)	(ST, CL, SC, CO, OF, AFQT)	(ST, CL, SC, CO, OF, AFQT)
R	.497	.369	.393	.342	.336	.415	.361
N	118	118	236	123	242	241	483
Z	1.19			.10	1.02		
	(All 11 Composites)	(All 11 Composites)	(All 11 Composites)	(All 11 Composites)	(All 11 Composites)	(All 11 Composites)	(All 11 Composites)
R					.374	.437	.378
N					242	241	483
Z					.81		

This multiple predictor variable set was obtained using PROC STEPWISE of the Statistical Analysis Software (SAS) with MODEL option MINB. As recommended in the SAS manual for this procedure, the model selected involved that set of predictor variables when the C_p statistic first approached the number of weights estimated—excluding the intercept. Correlations involving this predictor variable set were also obtained for ODD/EVEN halves using PROC ESQUAD for Test Forms 5-7 and 8-14 to provide added support for the validity of the obtained relationships (See Table 10). For relationships reported here, all ASVAB Scaled Scores and Identification performance had to be available for each soldier.

Z values were computed by the formula: $|Z_1 - Z_2| \sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}$. See McNemar, Q. Psychological Statistics, John Wiley and Sons, Inc. 1962.

pp. 139-140. For two-tailed tests, $|Z| > 1.96$ is significant at $p < .05$. Tabled Z s address the significance of difference between correlations obtained using independent halves of available data.

With N as the sample size and K as the number of predictor variables, Harrberg (1969) indicated that to obtain stable multiple correlations—across other samples—the N/K ratio should approach 20, i.e., there should be approximately 20 observations per weight estimated. Since ODD/EVEN halves contained approximately 120 observations each, the multiple correlations involving the best six ASVAB Scaled Scores for these Test Forms were also computed using PROC ESQUAD of SAS for the other Test Forms and all Test Forms to provide added support for the validity of the obtained relationships. See Harrberg, P. A. The parameters of cross-validation. Psychometrika Monograph Supplement, 1969. No. 16.

Table 3

Correlations Between Weighted/Unweighted ASVAB Scaled Scores Predictors and Vehicle Identification Performance

ASVAB Scaled Scores	Correlations for Total Sample					Weighted Correlations Based On Only	
	Unweighted	Weighted	Weighted	Using Weights Estimated From		ODD Half	KVKN Half
AFQT							
N	581	581	581	581	581	291	290
K (No. Wgts. Est.)	1	16	31	1	1	16	16
Z					.02		.04
CO							
N	583	583	583	583	583	292	291
K (No. Wgts. Est.)	1	16	31	1	1	16	16
Z					.06		1.21
FA							
N	583	583	583	583	583	292	291
K (No. Wgts. Est.)	1	16	31	1	1	16	16
Z					.32		.78
MM							
N	589	589	589	589	589	295	294
K (No. Wgts. Est.)	1	16	31	1	1	16	16
Z					0		.50
CM							
N	586	586	586	586	586	293	293
K (No. Wgts. Est.)	1	16	31	1	1	16	16
Z					.26		1.19
CL							
N	584	584	584	584	584	292	292
K (No. Wgts. Est.)	1	16	31	1	1	16	16
Z					.34		1.18
GT							
N	618	618	618	618	618	309	309
K (No. Wgts. Est.)	1	16	31	1	1	16	16
Z					.12		.07
EL							
N	583	583	583	583	583	292	291
K (No. Wgts. Est.)	1	16	31	1	1	16	16
Z					.04		.16
SC							
N	585	585	585	585	585	293	292
K (No. Wgts. Est.)	1	16	31	1	1	16	16
Z					.29		.69
ST							
N	586	586	586	586	586	293	293
K (No. Wgts. Est.)	1	16	31	1	1	16	16
Z					.19		.55
OF							
N	589	589	589	589	589	295	294
K (No. Wgts. Est.)	1	16	31	1	1	16	16
Z					.18		1.67

a,b superscripts for these correlations which are the same indicate no significant differences -- $p > .05$; where superscripts are different -- $p < .05$.

c Only this correlation failed to attain statistical significance ($p > .05$).

* $p < .05$; ** $p < .01$. All tests are two-tailed.

Table 4

Discriminant Analyses Using ASVAB Scaled Scores Test Forms 5-7/8-14 for a Calibration (ODD HALF), Test (EVEN HALF) and Total Sample

ASVAB Composites (Test Forms 5-7) Classified by a Linear Discriminant Function as					ASVAB Composites (Test Forms 8-14) Classified by a Linear Discriminant Function as				
		HIGH	LOW	TOTAL			HIGH	LOW	TOTAL
ODD HALF (CALIBRATION DATA)	APRIORI ACHIEVE GROUPS	HIGH	69	9	78	HIGH	74	9	83
			88.46	11.54	100.00		89.16	10.84	100.00
		LOW	25	15	40	LOW	30	11	41
			62.50	37.50	100.00		73.17	26.83	100.00
		TOTAL	94	24	118	TOTAL	104	20	124
		79.66	20.34	100.00			83.87	16.13	100.00
		$\chi^2 = 11.00, p < .001$ "Mite" = 71%					$\chi^2 = 5.18, p < .03$ "Mite" = 69%		
ASVAB Composites (Test Forms 5-7) Classified by the Linear Discriminant Function of the Calibration Data					ASVAB Composites (Test Forms 8-14) Classified by the Linear Discriminant Function of Calibration Data				
		HIGH	LOW	TOTAL			HIGH	LOW	TOTAL
EVEN HALF (TEST DATA)	APRIORI ACHIEVE GROUPS	HIGH	66	13	79	HIGH	75	7	82
			83.54	16.46	100.00		91.46	8.54	100.00
		LOW	26	13	39	LOW	33	6	41
			66.67	33.33	100.00		80.49	19.51	100.00
		TOTAL	92	26	118	TOTAL	108	13	123
		77.97	22.03	100.00			87.80	12.20	100.00
		$\chi^2 = 4.33, p < .05$ "Mite" = 67%					$\chi^2 = 3.07, p > .10$ "Mite" = 67%		
		$F(1,1) = 2.34, p > .05$					$F(1,1) = 1.68, p > .05$		
ASVAB Composites (Test Forms 5-7) Classified by a Quadratic Discriminant Function as					ASVAB Composites (Test Forms 8-14) Classified by a Quadratic Discriminant Function as				
		HIGH	LOW	TOTAL			HIGH	LOW	TOTAL
TOTAL SAMPLE	APRIORI ACHIEVE GROUPS	HIGH	132	25	157	HIGH	145	20	165
			84.08	15.92	100.00		87.88	12.12	100.00
		LOW	33	46	79	LOW	36	46	82
			41.77	58.23	100.00		43.90	56.10	100.00
		TOTAL	165	71	236	TOTAL	181	66	247
		69.92	30.08	100.00			73.28	26.72	100.00
		"Mite" = 75%					"Mite" = 77%		

Note 1: ACHIEVE groups were defined based on rank order of soldiers identification performance score. Those falling in the lower third were defined as LOW. PRIORS parameter in the Statistical Analysis Software (SAS) PROC DISCRIM was defined to reflect this definition.

Note 2: Tabled F is the ratio of the χ^2 (ODD) / χ^2 (EVEN). Generally the ratio of two independent chi-squares divided by their respective degrees of freedom (n_1, n_2) is defined as an F . See McNamee, Q. Psychological Statistics, John Wiley and Sons, Inc, 1962, pp. 230-231.

Note 3: For cases where results are presented for a quadratic discriminant function, the within covariance matrices were not homogeneous; therefore, within covariance matrices were used rather than the pooled covariance matrix.

Conclusions

- o Unweighted correlations are in the low .30s.
- o Differential weighting of ASVAB predictor categories led to modest but generally non-significant increases in correlations.
- o Multiple correlations were approximately of the same magnitude as correlations obtained by differentially weighting individual ASVAB predictors.
- o Quadratic discriminant functions involving ASVAB Scaled Scores were able to discriminate between high and low performing soldiers about 75% of the time.
- o High criterion reliability and supplementary analyses involving random sample halves generally point to the validity of relationships reported.
- o It would be desirable to develop new ASVAB Composites which better predict a wider array of Army skills.

New Assessment for Short-Service Volunteers

Axel R. Kaiser

Selection Center for Volunteers, Düsseldorf, FRG

The personnel resources of the armed forces in Germany will be dramatically reduced in the nineties. The use of hormonal ovulation inhibitors in the late sixties reduced the birth rate to one half within one decade. So the 1975 age class to be drafted into military service in 1994 is just 10 % more than the total number which is required by the armed forces (Figure 1). To meet the quantitative demands of the armed forces the service period will be extended to 18 months and medical classification criterions will be changed. The qualitative demands will become a special problem because of the requirements of business and industry in the nineties. It is not any more possible for the armed forces to pick out the elite like in most cases before. Efforts therefore have to shift from selection to placement of volunteers, carefully taking in account their developmental capacities and their deficiencies.

One of the steps to meet these demands is a reorganization of the current assessment system for volunteers. The new assessment for volunteers as well as a revised assessment for officer candidates (Melter, 1987) aim at a more efficient use of the professional skills and experiences of psychologists and officers by separating categories, observation and rating.

In the *current assessment system* (Figure 2) a board of one or two officers and one psychologist examine jointly the applicant, using the same set of categories which are rated on a scale from 1 to 9.

The categories which relate to personality are:

- 1 appearance
- 2 readiness for action, drive
- 3 decision making ability
- 4 stress resistance
- 5 sense of responsibility
- 6 social behavior.

Mental aptitude factors are:

- 7 apprehension
- 8 verbal expression
- 9 reasoning and judgement
- 10 planning and organization
- 11 learning motivation

Physical fitness and sports performance:

- 12 stress resistance in sports and physical fitness.

Figure 1

Reduced Personnel Resources (male)
of the Armed Forces in Germany

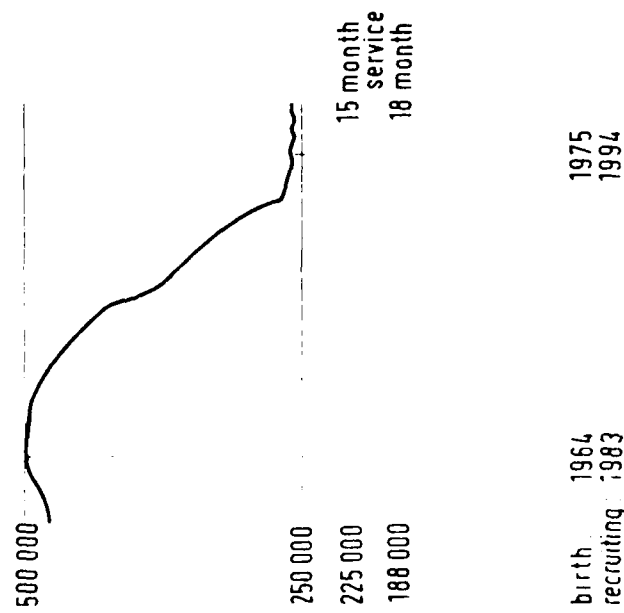
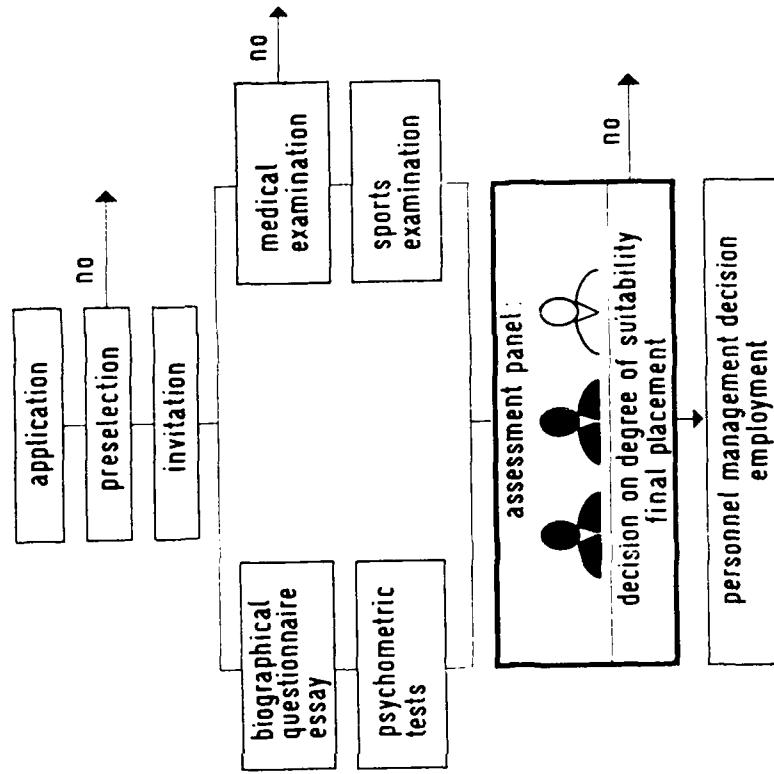


Figure 2

Current Assessment Procedure
for Short-Service Volunteers



In the new assessment the following set of categories for rating will be used:

		ratings by	
		officers	psychologists
1	leadership potential	x	x
2	conscientiousness	x	x
3	social competence/ cooperation	x	x
4	verbal expression	x	x
5	judgement and decisiveness	x	x
6	stress resistance		x
7	intelligence		x
8	achievement motivation		x
9	profession/ career orientation	x	
10	practical performance	x	
11	physical fitness/ stress resistance	x	

The rating scale is reduced from 9 to 7 points, all categories got behavior-focussed descriptions. Categories 6, 7 and 8 refer to psychological constructs.

The new assessment procedure emphasizes a separate collecting and rating of informations by psychologists and officers taking in account their different professional background (Figure 3).

Psychologists use the following instruments: psychometric tests, essay tests, a biographical questionnaire and the psychological interview. The psychological interview is considered to be a main resource for collecting, checking and combining data. There is sufficient experience that separate biographical or test data do not relate directly to complex assessment dimensions. They can only be seen as indicators for the overlapping and net-like structured characteristics of a person. Considering this holistic approach to personality nevertheless the psychological interview is focussed on behavior and guided by the principles of behavioral analysis. Complex findings in terms of functional descriptions - like compensative tendencies, avoidance behavior, attribution style and so on - as well as the prognosis of developmental capacities cannot be pressed into separately given assessment dimensions and do need therefore an additional report which contains also informations for training and future placement. The psychological interview needs a relaxed, non-examinative setting which gives way to a confident and open communication. This does not only meet the demands of a modern conception of partnership and cooperation between applicant and personnel assessor but is also appropriate to improve the reliability of informations. Especially non-directive accepting verbalization of the emotion-centered present state of the person supports decrease of overexcitement and facilitates a more relaxed and realistic communication. A final feed-back and counseling should be an element of any assessment interview to meet the personal needs of the applicant (Figure 4).

Figure 3

New Assessment Procedure for Short-Service Volunteers

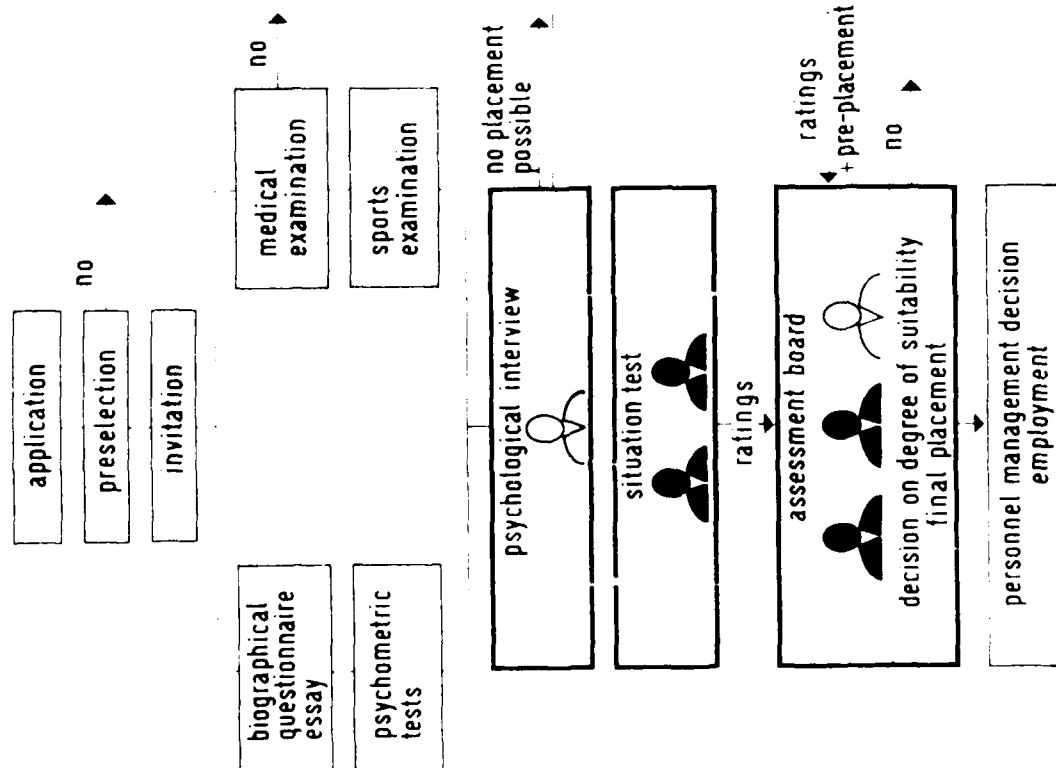
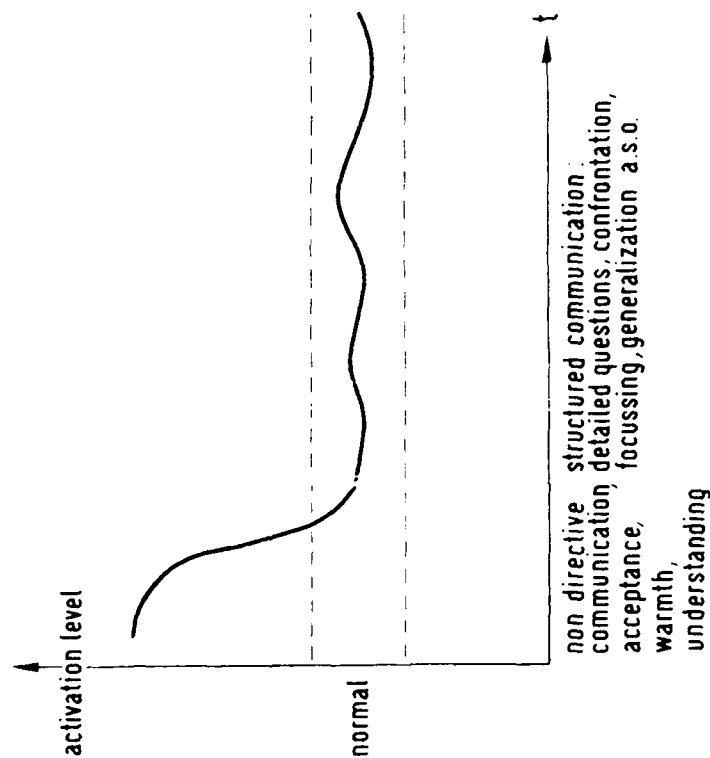


Figure 4

Communication Style and Activation Level



Board officers are in charge of a situation test which is followed by a short interview referring to the behavior within this test and to the military job orientation of the applicant. Two different kinds of situation tests are applied depending on the education level of the group of applicants being observed. The one is a complex planning task requiring mainly intellectual and organizational abilities, the other one emphasizes practical skills. Both situations provide observation facilities referring to factors like practical and organizational performance, cooperative behavior and the potential for non-commissioned officer's career. Officers are trained by psychologists in observing and rating different aspects of overt behavior.

The independently gained ratings of psychologists and officers are combined to an overall rating of the applicant including a rating of his development potential. Only in cases of disagreement there will be a board discussion. The assessment board finally makes a decision on the degree of suitability and placement. First organizational checks with the new assessment system have been done demonstrating that it can be practiced and is accepted by the personnel staff.

The internal reorganization of the current assessment system is accompanied by the development and application of computer assisted, computer adapted and simulation based testing. Considering the qualitative demands of the nineties this will not be sufficient.

For historical reasons psychologists are exclusively civilians and assessment procedures almost entirely are applied before men enter the Federal Armed Forces. So prognosis and placement have to cover an unsuitable long period regarding the multiple interactions between skills, training, leadership effects, job satisfaction and other factors within the forces. Applied military psychology should therefore care as well for educational and training support, additional sequential assessment procedures, for support concerning social psychological aspects and treatment of individual peculiarities. Neither the total personnel strength of the Psychological Service of the Federal Armed Forces nor the organizational structure provide the conditions for realizing that conception of an integrated military psychology (Mademann & Puzicha, 1987).

References

- Mademann, P.W. & Puzicha, K.J. (1987). *Revision of the Psychological Service of the Federal Armed Forces*. Proceedings of the 28th Annual Conference of the Military Testing Association, Mystic.
- Melter, A.H. (1987). *New Behavioral Assessment of Officer Candidates*. Proceedings of the 29th Annual Conference of the Military Testing Association, Ottawa.

METHODICAL AND ORGANIZATIONAL DEVELOPMENT OF SELECTION: PROGRESS AND RESULTS*

Albert H. Melter
Central Personnel Office
of the German Federal Armed Forces, Köln, FRG

Within the framework of revising the methodology of officer candidate selection procedures (STEEGE 1988, OTTE 1988) new aptitude ratings and test stations were conceptionally outlined (MELTER 1987). In 1988, aptitude factors and corresponding scoring procedures, situation tests, interviews, and questionnaires had been subjected to methodical tests and organizational checks. In addition, new test modules for the CAT software production have been prepared (UHLMANN 1988).

Purpose of Field Studies

We started gathering experience, whether competence-orientated allocation of responsibilities, improved preparation for biographical interviews, and changed systems of aptitude ratings and observation tools would result in better rating distributions.

Up to now, there was an strong fixing on the degree of restricted suitability and on decades of stable failure-rates. On average, 50% of applicants were non-suited, and most of the accepted had been signed as restricted. What factors substantiate these figures? Is it possible to organize the different contributions of military raters and of psychological competence in a more effective manner?

Method

Four boards - each with one psychologist and two officers - had been engaged in observation, judgement, and rating of performances and behaviors in partly separated test stations. Psychologists had to rate another set of aptitude factors than both officers. 1013 officer candidates had been tested over a period of January 17th to September 2nd, 1988.

Within phase A from January to April, board members separately prepared interviews with applicants' questionnaire responses and jointly carried out the board interview. The psychologist elucidated the applicant's biography over 20 minutes, followed by both officers who examined over 10 minutes profession and career orientations, objectives, interests, and behaviors in organizations. Then, officers were in charge of a situation test consisting of a complex planning task with the opportunity of specific interactions. The psychologist took part to rate his set of aptitude factors.

* The views expressed in this paper are those of the author and do not necessarily reflect official policies or positions.

Within phase B from May to August, interviews were separately prepared and separately carried out by the board members. The questionnaires had been revised with experiences from phase A. The psychologist had a minimum of 30 minutes for interviewing, officers a maximum of 20 minutes. First in phase B, interviews of board officers followed the situation test, which was administered without the psychologist. Later in phase B, interviewing had been given priority to situation test with a time-budget of 20 minutes. In this case, psychologist was not allowed to overrun 30 minutes for interviewing.

The complex planning tasks had been revised with experiences out of phase A, too. 75 minutes were available for groups of not more than six applicants to cope one out of seven planning tasks:

- setting up a sport shop
- planning a rescue exercise in a school
- organizing a holiday course, a young peoples' meeting place, a school prize-giving day, a sales department, a municipal area.

Written draft, short lecture, and group discussion had been outlined as segments of action and of observation. Officers scored segment- and overall-ratings of assigned aptitude factors.

In phase A, additional test stations were available for borderline cases. In B, there was only one additional test station "short lesson", in which psychologist and both officers served as raters. Additional test stations were held at third examination day.

Data Collection

Applicants completed the performance tests of intelligence, capacity of attention, and knowledge. The latter was limited to history, geography, politics, economy, technology, physics, and chemistry. The biographical questionnaire consisted of open and closed response-formated questions about biography, self-description and self-rating, profession and career orientation, and sport.

Self-description included 200 items assigned to 13 constructs such as achievement, affiliation, endurance. Only the psychologist received such data to prepare the interview. Additionally, he judged and rated an essay, behaviors in the situation test (only in phase A), and in the short lesson (phase B).

Officers judged and rated selected aptitude factors after the interview and during the situation and sport tests. After all test stations, officers and psychologist had to rate separately each aptitude factor of their set with 7-points overall-ratings and the degree of suitability categorized as good, suited, restricted, and non-suited for each applicant.

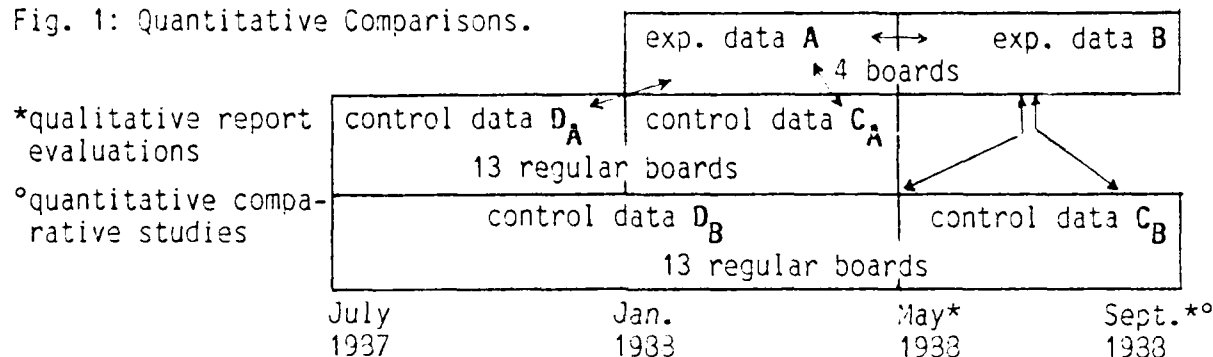
In phase A, boards determined now obvious positive or negative decisions and borderline cases at the end of the second examination day for further investigations at the third examination day. Then, final board discussions took place with a board-rating for five aptitude factors, adding psycholo-

gist's ratings and officers' ratings (without sport-rating), and determining the degree of suitability. This degree resulted from comparing the sum of 11 ratings with cut off points. Deviations from the point cutting off "restricted" from "non-suited" were exceptionally allowed for single cases, which had been assigned at least one poor (=7) rating in phase A or, in phase B, one poor rating from the psychologist and one from both officers.

Analytic Procedures.

The four experimental boards and their superiors had to report their qualitative organizational evaluations at the end of phase A and B. Experimental and control data from the regular boards had been compared only using descriptive statistics such as frequencies, one-way variance analysis, and correlations. The self-description questionnaire has been analyzed for item and scale characteristics and inter-correlations.

Fig. 1: Quantitative Comparisons.



Results

One-way variance analysis. Both experimental groups and their assigned control groups were comparable for N's, means of final school marks, school marks for mathematics and for sport, and for test scores of capacity of attention. A and B were statistically not comparable for intelligence and mathematics test scores.

Tab. 1: Comparisons between A, B and Control Groups.

Group	Period	N	Intelligence		Math. Test
			Verbal	Numerical	
A	Jan.-April	514	54.3*	69.5**	100.3
B	May -Sept.	504	52.3	67.5	97.2***
C _A	Jan.-April	2112	52.9	68.3	100.4
D _A	July-Jan.	2143	52.3	67.0	99.5
C _B	May -Sept.	3043	52.5	67.5	93.9
D _B	July-April	6234	52.5	67.5	99.9

* With $p \leq .01$ greater than all column means; ** greater than C_A and D_A column means; *** lower than A and D_B means of column.

Frequency of degree of suitability. The main result of the new assessment procedures was a better rating distribution in different settings.

Tab. 2: %-Distribution of Degrees in Experimental and Control Groups.

Degree	A	C _A	D _A	B	C _B	D _B
non-suited due to cut off point	N: 378 11	1311 25	2642 25	355 20	1331 21	4197 24
non-suited due to 1 or 2 poor ratings	25	7	10	17	9	8
good	15	3	2	14	3	2
suited	37	34	32	34	32	33
restricted	10	31	30	15	35	32

Percentage of non-suited (due to cut-off) applicants significantly decreased from 25% in C_A and D_A to 11% in A. For that, percentage of non-suited due to one or two poor ratings increased from 10% in D_A to 25% in A. This result wasn't desired. It could be diminished in phase B by introduction of a rule that only two or more poor ratings (7 on a 1-7 scale) from each competence justify the degree "exceptionally non-suited". In both experimental groups, "good" was more frequent and "restricted" was less frequent than in assigned control groups.

Frequency of aptitude factor ratings. The rating distributions of new aptitude factors turned out better than rating distributions of the regular aptitude factors. There were fewer and lower differences between A and B than to each assigned control group. Though using different aptitude factor sets, psychologist and officers surprisingly judged in a very stable way over the different settings of test stations in A and in B. It's obvious that altered settings for interviewing and situation testing had less effectiveness than rules for observation and rating.

Boards in A and B rated more rigorously than each member. Trends had been observed to rate more rigorously in B than in A and most rigorously in each assigned control group. The control boards had been used their 9-points scale more restrictively than the experimental boards their 7-points scale, in which psychologists and officers assigned some "very good" (=1) ratings to most aptitude factors.

Contributions of military raters and of psychological competence. The effectiveness of each board member had been investigated by comparative frequency analysis of degree ratings recommended before entering board discussions.

Percentages of good-recommended applicants were greater in B than in A for all board members. The same applied to non-suited recommendations. Percentages of suited- and restricted-recommendations were lower in B than in A. It's possible that classifications of performances and behaviors in the test stations were more successful in B than in A (Tab. 3).

Tab. 3: Separate %-Distributions of Degrees of Suitability for Psychologists, Staff Officers, and Officers ($N_A=375$; $N_B=331$).

Degree	PSY		STO		OFC	
	A	B	A	B	A	B
good	6	8	5	7	5	6
suited	31	30	29	24	27	23
restricted	26	24	30	27	31	26
non-suited	36	38	36	42	38	44

Recommendations before entering board discussions were congruent in most cases, but more in A than in B, in which boards had been more challenged to lead differences between recommendations from each board member to the board's final degree of the applicant.

In both phases, psychologist and officers rated the five common aptitude factors in their specific set more rigorously than expected. The aptitude factors sport and reasoning, both more defined as performance measures than others, had the statistically best distributions.

Interview and situation test. More convenient conditions for interviewing correlated with increase of restricted-recommendations. Especially officers rated rigorously, if they had available moretime for interviewing. Including psychological competence into the situation test is desirable for methodical reasons. Observing, judging, and rating were statistically better in B than in A.

Biographical data. Score distributions of the self-description questionnaire indicated applicants' self-monitoring as behaving in a social desirable way. Only achievement-orientated questions such as endurance, learning, leadership, and achievement significantly correlated with at least five aptitude factors above all in B. This setting had supported psychologists' using questionnaire and essay informations for rating the aptitude factors before entering the board discussions.

Short lesson. Officers rated more rigorously than psychologists. The latter made always use of the whole range. But short lesson as test station can't be any longer realization of applicants' stress resistance. Correlations with relevant aptitude factors in psychologists' set are too low.

Discussion

In appreciation of the results one must take into consideration the fact that they come from a field study, whose quality of data allows only descriptive statistics. Objectives of the methodical and organizational development of personnel selection are unambiguous definition and responsibility of psychological competence, especially with respect to recommended course of studies at an Armed Forces university.

To meet such objectives, results suggest introducing the specific sets of

aptitude factors and the 7-points scale. In addition, opportunities for rater supervision and training have to be enlarged to overcome rigorous rating.

For the long term, one should abandon of aptitude factors such as career orientation, leadership potential, decisiveness, and planning in rating applicants without military training. Such factors could be tested in an assessment center for first-trained draftees, volunteers, or cadets in search of developing junior leaders. Applicants without military training should only complete psychological tests with respect to their basic social, verbal, numerical, and learning abilities and they should undergo the obligatory medical check of their state of health before allowing for employment.

The results also suggest the separate interviewing by psychologist and officers including biographical and specific self-descriptive data. Test and essay interpretation should exclusively be psychological affairs. The situation test with a complex planning task should be carried out by officers in the tested way but with the opportunity for the psychologist to rate social aptitudes in his factor set.

References

- Melter, A. H. (1987). New Behavioral Assessment of Officer Candidates. Proceedings of the 29th Annual Conference of the Military Testing Association, Ottawa.
- Otte, R. (1988). Zur Auswahl von Bewerbern für die Einstellung in die Laufbahn der Offiziere. Vortrag auf dem 3. Deutschen Assessment Center Kongreß, München.
- Steege, F. W. (1988). Prüfmethdische Weiterentwicklungen im Annahmeverfahren für Offizierbewerber. Vortrag auf der 3. Arbeitstagung zur Weiterentwicklung des Annahmeverfahrens für Offizierbewerber, Waldbröl.
- Uhlmann, L. (1988). Kognitive Fähigkeitstests als zum Assessment Center alternative Methoden. Vortrag auf dem 3. Deutschen Assessment Center Kongreß, München.

TESTING U.S. ARMY WAR COLLEGE STUDENTS' WRITING ABILITIES

Colonel Robert J. Davis and
Professor Jim Hanlon
U.S. Army War College

BACKGROUND

In 1980 the U.S. Army War College (USAWC) began contracting for supplemental instruction in writing for selected USAWC students. The course was first called Writing Enhancement and then renamed Writing Effectiveness. Students participated voluntarily. In 1984, as part of a program of self-assessment, students tested themselves in English mechanics and expression with the Missouri College English Test. Test results enabled the students to decide whether to take the course, which was offered in one-hour weekly sessions for some fourteen weeks. Typically, some 20-25 students would indicate initial interest in the course. The course focused on the grammar, mechanics, and usage of standard written English. However, only eight or so students would regularly attend throughout the course of instruction. The course instructor assisted a few students in developing drafts and polished final texts of their Military Studies Project (MSP), a more extensive capstone paper required of each USAWC student. The course began in October and extended through January.

USAWC COMPREHENSIVE WRITING PROGRAM (1987-1988)

In 1987 the USAWC committed itself to greater emphasis on writing in the curriculum. Coordinating its program with the Army Writing Office of the Training and Doctrine Command (TRADOC), USAWC upgraded writing instruction by focusing on (1) faculty development in writing, (2) more rigorous testing in writing, and (3) a mandatory course on instruction for selected students, supplemented by a longer voluntary course of intensive review of grammar, mechanics, and usage.

Faculty Development. In 1988 USAWC retained a consultant to provide on-going assistance in orienting new faculty as well as to work with older faculty in order to use writing more effectively in the curriculum. Consequently, faculty expectations for students' writings are now more precisely indicated in writing assignments. In addition, assignments themselves, through master planning, are spaced carefully throughout the academic calendar to insure that students have adequate planning and working time to concentrate on given assignments. An evaluation sheet has been developed to assure that all written work meets USAWC standards in the following areas: focus, critical reasoning, organization, use of source materials, readability, and presentability. Faculty are now more

cognizant of writing as a vehicle of instruction, rather than simply as a course or program requirement. Instruction focuses more on in-progress assessment; so revision of drafts, based on instructors' recommendations, is more commonplace. Further, instructors refer students who need closer assistance to the Effective Writing Course. The faculty development consultant, who has implemented an effective program of writing-across-the-curriculum at his home institution, offers workshops for new faculty and conducts follow-up activities for all faculty throughout the academic year.

Testing. All U.S. students, during their August orientation to USAWC, now take the Test of Standard Written English (TSWE) and submit a 45-minute impromptu writing sample on a common topic. TSWE provides an indirect measure of writing abilities. Developed in the past 15 years by the Educational Testing Service, this 30-minute, 50-question test indicates a person's ability to recognize errors in mechanics and usage and to sort out more desirable from less desirable stylistic options. The writing sample provides a direct measure or indication of writing abilities, albeit under severe constraints. It is holistically assessed twice on a 1-6 scale by experienced English professors--with a third reading in the event of divergent readings. Raters use a scoring rubric developed from close analysis of characteristics identified in a random sample of the group's writings. USAWC students self-score their TSWEs for on-the-spot feedback; and, through their faculty advisers, they receive a prompt assessment of their writing sample. Students whose TSWE scores exceed 15 incorrect responses and whose holistic assessments are on the low end of the rating scale (1-1, 1-2, and 2-2) are identified as candidates for the mandatory course. For the last two years, some 20-22 "hardcore" students have been selected for this course. Further, some 15-20 additional students have been strongly recommended to take the instruction.

Writing Courses. USAWC writing instruction is currently offered through three courses of instruction: (1) a ten-week (one hour weekly) course in Effective Writing, (2) an eighteen-week (one hour weekly) Intensive Review of Grammar, Mechanics, and Usage, and (3) a follow-up offering of individualized instruction on drafts of students' MSPs.

Some 40 students currently take Effective Writing, which emphasizes the Official Style. The course focuses on revising prose through Richard Lanham's Paramedic Method. It emphasizes the process of writing, rather than the product. The sessions concentrate on active and passive voice, avoidance of wordiness, using parallel structures, shifting from operational to deliberative discourse, using collaborative strategies for composing, using voice effectively in writing, and proofreading for polished final copy. The final session offers an overview and closing perspective on effective writing. Throughout the course, many students voluntarily submit drafts of current

writings for the instructor's review and feedback. The students are strongly encouraged to practice strategies of revision on their own writings. Effective Writing is offered from mid-September through November.

Approximately 14 students of the Effective Writing Course also participate in an Intensive Review of Grammar, Mechanics, and Usage. All students volunteer for this course, but they are committed to regular attendance after volunteering. The course was developed when a few students in 1987 noted that they would like a more thorough refresher on the basics of writing than was offered in Effective Writing. The course offers a close review of grammar for standard written English, error recognition, mechanics of punctuation, problems of usage, and sentence-level revision. Each meeting provides hands-on exercises which focus on given problems of grammar, mechanics, and usage. For the first ten weeks, the course runs concurrently with Effective Writing, but on a different day of the week. It continues thereafter weekly through February.

Some two-thirds of the students in both courses seek follow-up instruction through editorial advice and conference on drafts of their MSPs. This activity sometimes begins early in the academic year and continues until final drafts are prepared in March. All USAWC students are eligible for this instruction, but students of Effective Writing have first priority.

OBSERVATIONS AND OUTCOMES

1. Testing for writing abilities is integrated into the total course of instruction.

2. Indirect testing of writing abilities supports direct testing. Occasionally students test well indirectly, but their impromptu writings do not corroborate the direct testing. These students are generally recommended to take the Effective Writing Course.

3. Despite sustained efforts to assure validity and reliability, testing for writing abilities is necessarily arbitrary and occasionally fallible. If a student demonstrates capabilities not indicated by testing (usually through other writings), then the student is excused from the Effective Writing Course.

4. Writing skills can be honed and refined at any developmental level. In fact, mature professionals who have time to reflect and a good reason to write are ideal candidates for writing instruction.

5. Testing for writing abilities is grounded in the curricular and professional environment of USAWC. Topics for the writing samples call for deliberative, speculative, and values-oriented responses. Thus the testing clearly signals broad curricular objectives.

6. Testing for writing abilities enhances the role of writing in the curriculum. Offering the writing tests during orientation week provides a meaningful context for alerting the new students to the importance of writing and to the kinds of writing requirements and performance standards that are required. Thus testing establishes a high priority on the skill being tested.

7. Structured, formalized testing takes place as students enter the program of instruction. Thereafter, "testing" is no longer an issue. Rather, students are offered on-going support in meeting institutional requirements in writing. A few students of the Effective Writing Course have become published writers, and more than a few have submitted impressive MSPs. So testing does not purport to monitor or somehow facilitate "minimal competence." Rather it leads to non-threatening institutional support in writing and often to enhanced performance in the students' written work.

References

Cooper, Charles R. and Odell, Lee (Eds.). (1975) Evaluating Writing: Describing, Measuring, Judging. Urbana, IL: National Council of Teachers of English.

Lanham, Richard A. (1986) Revising Prose. 2d ed. New York: MacMillan.

Reinking, James A. et. al. (1981) Improving College Writing. New York: St. Martin.

White, Edward M. (1985) Teaching and Assessing Writing. San Francisco: Jossey-Bass.

Development of a New Pseudo AFQT to Detect Possible Compromise

Thomas W. Watson
Steven W. Hoffer, Sgt, USAF
Malcolm James Ree

Air Force Human Resources Laboratory

The Armed Forces Qualification Test (AFQT) is formed from selected subtests of the Armed Services Vocational Aptitude Battery (ASVAB). The AFQT is used by the United States military services as the primary means of determining qualification for military service. Since the subtests comprising the AFQT impact an applicant's prospects for military service, performance on them may be compromised. Further, it is assumed that if compromise occurs, it will be likely on the AFQT subtests of the ASVAB only since these are the subtests used to determine enlistment standards. To detect if compromise is likely, the services have developed a "Pseudo AFQT" composed of non-AFQT subtests from the ASVAB (on which compromise is unlikely) which are highly correlated with the AFQT and which are used to predict it. Large differences (called deltas) between predicted AFQT scores (i.e., the Pseudo-AFQT) and actual AFQT scores are used as an index of possible cheating. Historically, deltas larger than a selected cutoff resulted in verification retesting. Such retesting no longer takes place, but the United States Military Entrance Processing Command (MEPCOM) uses Pseudo scores and deltas for historical tracking of the likelihood of AFQT compromise.

A Pseudo AFQT has been used for many years and developmental work on an earlier version was conducted by the Center for Naval Analyses (Sims & Truss, 1982). More recently, Wegner & Ree (1986) of the Air Force Human Resources Laboratory (AFHRL) examined the relative merit of a variety of alternative Pseudo-AFQTs. In the past, speeded subtests were included in the computation of the AFQT and the Pseudo AFQT. However, recently the Joint Services Selection and Classification Working Group (JSSCWG) decided to include only power subtests in computing both the AFQT and the Pseudo AFQT. Therefore, AFHRL was tasked with (1) developing a new Pseudo AFQT against the new AFQT criterion, (2) developing regression weighted adjustments to reduce bias vis-a-vis selected subgroups, and (3) computing deltas and delta cutoffs. The purpose of this paper is to describe and document that process.

Method

Subjects. This research used ASVAB scores and demographic data collected from young adults by the National Opinion Research Center (NORC) in their 1980 Profile of American Youth study (McWilliams, 1980). The sample contained 9,172 youth, ages 18 through 23 (4,550 males and 4,622 females), and was statistically weighted to be representative of approximately 25 million 1980 American youth. This same sample was used in the recent renorming of the ASVAB. The sample was also "clean" in that there was no motivation for compromise. This is due to the fact that ASVAB scores from this data base could not be used operationally for entry into military service.

Pseudo construction and alternative model testing. Using standardized ASVAB subtest scores and demographic information from this data base, a variety of Pseudo AFQTs were constructed, with and without consideration of subgroup membership. Full versus restricted and stepwise regression models were compared to determine (1) the magnitude of association between selected non-AFQT subtests and the new AFQT and (2) which, if any, demographic factors enhanced predictive efficiency in a practical way. Practical significance was defined as the ability of a single predictor to increment R^2 by at least .02 beyond the contribution of other predictors. The basic computation of the Pseudo consisted of the best weighted combination of two of the most promising non-AFQT subtests identified by Wegner and Ree (1986), regressed against the new computation of the AFQT. Other models (i.e., alternative Pseudo-AFQTs) used the same new AFQT

criterion and included the same subtests as in the basic Pseudo computation. However, in each of these alternative models, one or more of the following demographic variables were added to the predictor set: Gender, Ethnic Group (black, white, hispanic, other), and Education. Education was first defined in terms of highest grade completed (9, 10, 11, 12, 13+) and whether a subject had completed high school or obtained his/her GED. As will be discussed in greater detail below, Education was later redefined in terms of college versus no college. In addition to generating basic and alternative Pseudo regression models, means and standard deviations of the basic Pseudo AFQT were computed for the total sample and selected subgroups.

Computation of deltas and determination of false positive cutoffs. Once the basic and alternative pseudos were computed, adjusted and unadjusted deltas (difference scores) were calculated for each subject. Unadjusted deltas were computed by subtracting a subject's basic Pseudo AFQT standard score from his/her AFQT standard score. It should be noted that unadjusted deltas were anticipated to be of little importance except to indicate the extent to which subgroup adjustments were needed. Adjusted deltas were computed by subtracting a best-weighted adjustment from the unadjusted delta. Means and standard deviations were also calculated for both the unadjusted and adjusted deltas for the total sample and selected subgroups. Subgroup statistics were computed for the following groups: Gender, Education (college versus no college), AFQT Category¹, and Ethnic Group by Gender. The selection of factors to include in the adjustment evolved in an iterative fashion as mean adjusted deltas for various subgroups, computed from three different algorithms, were examined. The objective was to have mean adjusted deltas approach zero, not only for the total sample, but for all subgroups. Once a suitable delta adjustment was determined, false positive adjusted delta cutoffs were identified at the one-, two-, and five percent levels. For instance, at the one percent level, 99% of all subjects had adjusted deltas below the cutoff value. The term false positive is used since cheating was not expected in the data set used.

Results and Conclusions

Pseudo AFQT descriptive statistics. Means and standard deviations for the basic Pseudo AFQT are presented in Table 1. The Pseudo AFQT is the sum of the products of the regression weights and subtest standard scores and the regression constant for Subtests 1 and 2. The specific computation of the basic Pseudo AFQT, using the weights from the regression analyses discussed below, is as follows: $\text{Pseudo AFQT} = 39.1964 + (2.4921 * \text{Subtest 1}^2) + (.7275 * \text{Subtest 2})$.

Table 1.
Means and Standard Deviations for the Basic Pseudo AFQT

	Mean	Standard Deviation
Total Sample	199.94	30.49
Male	207.11	31.87
Female	192.55	27.08
Has Attended College	214.56	26.55
Has Not Attended College	191.26	29.33
Male & Has Attended College	224.52	25.49
Male & Has Not Attended College	197.72	31.00
Female & Has Attended College	205.46	24.12
Female & Has Not Attended College	184.12	25.54

¹ Although sometimes defined more specifically, AFQT Category used in this research is a six-category index of general trainability based on percentile. I = 93 - 100, II = 65 - 92, IIIA = 50 - 64, IIIB = 31 - 49, IV = 10 - 30, V = 1 - 9.

² To limit compromise, the Pseudo AFQT subtests will not be identified by name. Rather, they will be called Subtest 1 and Subtest 2.

As indicated in Table 1, means ranged from 184.12 for females who had not attended college to 224.52 for males who had attended college. Results follow expected patterns with males scoring better than females and college educated subjects scoring better than those with no college. Standard deviations were quite consistent across groups and ranged from 25.54 for females with no college to 31.87 for males in general.

Regression results. Results of the regression analyses are provided in Table 2. Full versus restricted model comparisons indicated that demographic factors (Gender, Education, and Ethnic Group) contributed little to prediction beyond the contribution made by the two subtests identified by Wegner and Kee (1982). The two subtests alone were able account for 70% of the variance in the AFQT criterion. When the demographic variables were added as predictors, they collectively enhanced explained variance by less than 7%. In stepwise analyses, the correlation between the first subtest and the criterion was almost as great as the combined correlation between the two subtests and the AFQT. In addition, college attendance (high school graduation or GED and 13+ years of education) was the only potent Education variable, entering the equation second, and Gender was the only other demographic variable to make a practical contribution to the increment in explained variance. The two subtests and the two demographic variables accounted for close to 76% of the criterion variance, while all other predictors in combination contributed only slightly more than 1% to explained variance.

Table 2.
Results of Regression Analyses

Full vs Restricted Model Results:				
Predictors	Multiple R	R ²	R ² Change	
Subtests 1 & 2	.839	.705		
Subtests 1 & 2 + Gender + Educat. + Ethnic Group	.879	.772	.067	
Stepwise Results:				
Step	Variable	Multiple R	R ²	R ² Change
1	Subtest 1	.827	.683	
2	College	.844	.713	.030
3	Subtest 2	.857	.734	.021
4	Gender	.871	.759	.015
5 - 12	All others	.879	.772	.013

On the basis of the regression analyses, the investigators concluded that the basic Pseudo AFQT, composed only of subtests 1 & 2, would be suitable³, and that adjustments which might be necessary could be made to delta.

With regard to the demographic variables, the regression results suggest that demographics contribute little, and that Education need not be broken down by year since only college experience had a practical but slight impact on prediction. In addition, Gender was the only other demographic variable to have a similar impact. The regression results suggest further that Ethnic Group membership makes no practical contribution to prediction. As a result, the

³ The Pseudo AFQT could have been calculated using only Subtest 1 since it correlated almost as well with the criterion as did Subtests 1 and 2 in combination. However, the investigators and members of the JSSCWG decided that use of both subtests was preferable to make compromise more difficult.

investigators aggregated Education for further analyses in terms of a dichotomous "college versus no college" factor and used only Education and Gender in initial adjustments to delta. As discussed below, AFQT score and an Ethnic Group * Gender interaction variable were added later in computing subsequent delta adjustments.

Mean Deltas and Adjustments to Delta. Examination of mean unadjusted deltas allowed the investigators to determine if subgroup adjustments were necessary (i.e., if they deviated substantially from zero). Likewise, examination of mean adjusted deltas based on different combinations of weighted predictors allowed the investigators to refine adjustments in an interactive fashion until mean adjusted deltas approached zero for all subgroups.

Mean unadjusted deltas for the total sample and for selected subgroups are presented in Table 3. Although mean unadjusted delta for the total sample was zero, mean unadjusted deltas for subgroups deviated appreciably from zero, ranging from -10.72 to 12.74. This indicated that adjustment was desirable. On the basis of the regression results summarized earlier, Gender and Education were regressed on unadjusted delta to determine an appropriate constant and least squares weights for an initial adjustment. This adjustment, called Adjustment 1, was computed as follows and subtracted from unadjusted delta: Adjustment 1 = $1.28 - (11.19 * \text{Gender}) + (11.81 * \text{Education})$.

Table 3.
Unadjusted Delta Means and Standard Deviations

	Mean	Standard Deviation
Total Sample	00.00	19.73
Male	-05.77	18.62
Female	05.94	19.08
College	07.75	17.33
No College	-04.59	19.64
Male, College	02.28	15.96
Male, No College	-10.12	18.50
Female, College	12.74	17.02
Female, No College	01.51	19.04

As indicated in Table 4, Adjustment 1 moved mean deltas toward zero when the subgroups in question were the same as in Table 3. However, as shown in Table 5, when mean adjusted deltas within AFQT categories were examined, Adjustment 1 was less successful. As a result, the investigators added AFQT in standard score metric to the adjustment predictor set and regressed Gender, Education and AFQT on unadjusted Delta to determine an appropriate constant and least-squares weights for a second adjustment. This adjustment, called Adjustment 2, was computed as follows and subtracted from unadjusted delta: Adjustment 2 = $-52.92 - (12.5 * \text{Gender}) + (1.27 * \text{Education}) + (.29 * \text{AFQT})$.

Table 4.
Adjusted Delta Means and Standard Deviations

	Mean	Standard Deviation
Total Sample	00.00	15.32
Male	00.00	15.76
Female	00.00	14.83
College	00.00	15.00
No College	00.00	15.50
Male, College	-00.26	15.04
Male, No College	00.14	16.15
Female, College	00.23	14.97
Female, No College	00.15	14.75

Table 5.
Mean Adjusted Deltas by AFQT Category

AFQT Cat.	Total Sample	Male	Female	College	No Col.	Male Col.	Male No Col.	Female Col.	Female No Col.
With Adjustment for Gender & Education Only (Adjustment 1)									
I	7.32	6.01	9.62	5.18	19.15	3.92	17.40	7.39	22.35
II	8.24	7.30	9.29	4.06	14.36	3.45	17.06	4.64	17.58
IIIA	3.87	3.12	4.55	-2.69	7.85	-3.10	5.80	-2.43	10.22
IIIB	-.15	-.91	.45	-5.93	1.58	-5.42	.22	-6.24	2.74
IV	-9.40	-8.59	-10.19	-15.19	-8.24	-15.04	-7.67	-15.77	-8.84
V	-16.07	-13.04	-19.69	-26.11	-15.64	-20.71	-12.78	-29.98	-19.12
With Adjustment for Gender, Education & AFQT (Adjustment 2)									
I	-2.79	-3.65	-1.28	-3.30	.06	-4.08	-1.27	-1.94	2.50
II	.36	.10	1.92	.55	1.55	.27	-.10	.83	3.65
IIIA	.57	-.34	1.41	.29	.75	.66	-.77	.06	2.50
IIIB	.26	-.26	.68	2.26	-.33	3.53	-1.21	1.46	.42
IV	-1.27	-.14	-2.38	.64	-.64	1.99	-.44	-.19	-2.91
V	.90	4.55	-3.45	-1.09	.99	5.74	4.51	-5.99	-3.31

As can be seen from the second data set in Table 5, Adjustment 2 was moderately successful in moving mean deltas toward zero in most AFQT categories. In those instances where mean adjusted delta had an absolute value in excess of two, it was for subjects in the highest AFQT

Table 6
Adjusted Delta Means and Standard Deviations, Gender by Ethnic Group

	Mean	Standard Deviation
Based on Adjustment 2		
White Male	-1.53	15.85
White Female	-0.41	14.58
Black Male	7.32	15.54
Black Female	-1.11	14.41
Hispanic Male	7.66	16.63
Hispanic Female	1.36	14.31
Based on Adjustment 1		
White Male	1.11	14.11
White Female	1.11	14.11
Black Male	1.11	14.11
Black Female	1.11	14.11
Hispanic Male	1.11	14.11
Hispanic Female	1.11	14.11

categories, who would have less motivation to compromise the AFQT, or for those in the lowest categories, who would not be eligible for enlistment.

The investigators were initially pleased with the second adjustment. However, on the advice of Dr Divgi of the Center for Naval Analyses (CNA), Adjustment 2 was examined for the sample partitioned by Gender within Ethnic Group. As indicated in Table 6, Adjustment 2 did not move mean deltas for these groups toward zero as well as it had for the groups in Table 4. Thus, the investigators added a Gender * Ethnic Group interaction term to the adjustment predictor set and regressed Gender, Education, AFQT, and Gender * Ethnic Group on unadjusted delta to determine the appropriate constant and least-squares weights for a third Adjustment. In addition, Pseudo scores rather than an AFQT scores were included as adjustment factors since they are less susceptible to inflation through cheating in operational use. Also, at the request of MEPCOM, the Pseudo in percentile form was used. As is indicated in Table 6, Adjustment 3 was more successful than Adjustment 2 in moving mean adjusted deltas toward zero for the Gender within Ethnic Group samples. Adjustment 3 is therefore considered the best adjustment and was recommended for operational use by MEPCOM. The computation of Adjustment 3, in detail, is as follows: Adjustment 3 = 4.0335 (Constant) + $(12.1322 * 1 \text{ if attended college; } 0 \text{ otherwise}) - (.0830 * \text{Pseudo AFQT percentile score}) - (8.5772 * 1 \text{ if male gender; } 0 \text{ otherwise}) - (3.7880 * 1 \text{ if hispanic female; } 0 \text{ otherwise}) + (1.2549 * 1 \text{ if hispanic male; } 0 \text{ otherwise}) - (7.4982 * 1 \text{ if black female; } 0 \text{ otherwise}) - (1.4293 * 1 \text{ if black male; } 0 \text{ otherwise}) + (.9196 * 1 \text{ if white female; } 0 \text{ otherwise}) + (1.4423 * 1 \text{ if white male; } 0 \text{ otherwise})$.

One- Two- and Five-Percent Adjusted Delta Cutoffs. Adjusted delta cutoffs using Adjustment 3, for the total sample, are provided below at the 1%, 2%, and 5% levels. These are adjusted delta values in a percentile score metric, greater than those attained by 99%, 98% and 95% of the weighted NORC sample respectively. They serve as practical decision points for the assumption of compromise. They are as follows: 5%: 24; 2%: 31; and 1%: 35. Note that they are false positive cutoffs since actual compromise in the NORC data base is improbable. It should also be noted that even when used on applicant data in an operational setting, cutoffs do not provide conclusive evidence of cheating; rather, they indicate a situation where compromise is highly probable.

References

- McWilliams, H. A. (1980). Profile of American youth: Field report. Chicago, IL: National Opinion Research Center.
- Sims, W. H., & Truss, A. R. (1982). The development and application of a pseudo AFQT for ASVAB 8/9/10. (CRC 470). Alexandria VA: Center for Naval Analyses.
- Wegman, I. G., & Ree, M. J. (1986, September). Alternative Armed Forces Qualification Test composites. (AFHRL-IP-86-27; AD-A 175 027). San Antonio, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Pre-operational Validation of New Army Flight Aptitude Selection Tests

D. Michael McAnulty
Anacapa Sciences, Inc.¹

Background

The U.S. Army conducts an Initial Entry Rotary Wing (IERW) training program at Fort Rucker, Alabama, for both commissioned officer (CO) and warrant officer candidate (WOC) helicopter flight students. The IERW course lasts approximately 9 months and is divided into primary, instrument, transition (if required), basic combat skills, and night flight training phases. Student performance is evaluated by daily flight grades, academic examinations, and end-of-phase flight checkrides. Upon completion of the IERW course, the students are assigned to operational units for further individual, crew, and team training.

The Army Research Institute Aviation Research and Development Activity at Fort Rucker has a continuing requirement to evaluate and improve the tests that are used to select IERW applicants. The original Flight Aptitude Selection Test (FAST) was implemented in 1966 in response to the unacceptably high attrition rates in the flight training program during the 1950s (Kaplan, 1965). The two FAST batteries, one for CO applicants and one for WOC applicants, resulted in a substantial reduction in the flight training attrition rates.

Subsequently, the FAST was revised to produce a single battery with fewer, shorter, and more reliably scored subtests (Eastman & McMullen, 1978). Shortened versions of 7 of the 12 FAST subtests were retained in the revised FAST (REFAST). The REFAST, implemented in 1980, is approximately one-half the length of the FAST. In 1984, Lockwood and Shipley found that the correlation between the REFAST and IERW performance was statistically significant. However, some of the subtests had undesirable psychometric characteristics and the REFAST accounted for only a small percentage of the variance in IERW performance. They concluded that a new FAST (NEFAST) battery was needed to improve the reliability and validity of the IERW selection process and to provide an equivalent term for rotating applicants.

At Fort Rucker, Alabama. In the first phase of the NEFAST development, research was conducted to determine the

¹This research was conducted as part of the Army Research Institute Aviation Research and Development Activity at Fort Rucker, Alabama. The author is grateful to the staff of the Fort Rucker Army Research Institute for their assistance and support. The research was conducted under the supervision of the Army Research Institute Aviation Research and Development Activity at Fort Rucker, Alabama.

ability requirements for the successful completion of IERW training. Experienced IERW instructor pilots were asked (a) to identify the tasks that are most indicative of successful performance in the primary and instrument phases of IERW and (b) to rate the type and importance of the abilities that are required to perform each task. The instructor pilot ratings were then transformed to a normally distributed, equal-interval scale using the method of successive intervals. Analyses of the ratings indicated that 24 abilities in the psychomotor, perceptual, language, and cognitive domains were required for successful performance in IERW. Twelve of these abilities were selected for test development on the basis of their amenability to a paper-and-pencil format and their measurement potential (McAnulty, Jones, Cohen, & Lockwood, 1984).

Experimental NFAST Development. Nine new tests were developed for an experimental NFAST battery. Eight tests were each designed to measure a unique ability and one test was designed to measure a complex of abilities required for the successful completion of IERW training. The complex test was designed primarily to measure decision-making ability, but it also assesses several other cognitive and perceptual abilities. Four standardized tests were also included in the battery as marker variables. The 7-hour experimental battery was administered to 273 general population subjects at three Army installations. The results indicated that the complex ability test and six of the unique ability tests assess reliable individual differences in the abilities of interest. The remaining two unique ability tests had undesirable psychometric characteristics or did not contribute any unique variance to the factor structure of the battery (McAnulty, Cross, & Jones, 1986). The experimental research results were used to produce two alternate forms of the NFAST.

Research Objectives. The purpose of this research is to evaluate the two alternate NFAST forms before implementing the battery to select applicants for helicopter pilot training. Specifically, the pre-operational validation research will be conducted (a) to determine the relationship between the NFAST tests, other predictor data, and performance in IERW training and (b) to equate the alternate forms of the battery on a large sample of flight students. The results can then be used to produce the two operational NFAST batteries.

Method

This research employed a predictive validity paradigm to evaluate the NFAST. However, the research subjects were previously selected for IERW training using the current selection procedures.

NFAST Battery. Form 1 of the NFAST validation battery consists of six subtests and the complex test and the

unique ability tests that had previously exhibited acceptable psychometric characteristics (see Table 1). Each of the new tests are divided into two or more sections. The sections of the Finding Rules and Sound Reasoning tests are designed to be equivalent. The sections of the other tests represent different approaches to measuring the ability construct or vary in the levels of difficulty. For example, the complex Flight Planning test is divided into six sections and three levels of difficulty. The levels of difficulty are defined in terms of the number of rules that must be memorized and applied to increasingly complex route maps.

Table 1

Composition of the Alternate Forms of the NFAST

Test	Ability	Sections	Items	Minutes
Flight Planning	Decision Making	6	72	48
Chart Use	Information Ordering	4	24	20
Figure Orientation	Spatial Orientation	2	56	6
Finding Figures	Closure Flexibility	2	80	6
Finding Rules	Inductive Reasoning	2	20	8
Rapid Match	Perceptual Speed	2	64	4
Sound Reasoning	Deductive Reasoning	2	20	6
Helicopter Knowledge	Knowledge (RFAST)	1	20	10

The validation battery tests are approximately two-thirds the length of the experimental battery tests. The Flight Planning and Chart Use test forms do not have any identical items. The alternate forms of the remaining five ability tests have approximately 50% of the items in common. Finally, a knowledge test of helicopter operating principles was adapted from the RFAST for inclusion in the validation battery. The items on the knowledge test are identical on both forms.

Procedures. For a period of 6 months, the NFAST batteries were administered to each entering CO and WCO class during the first week of IEKW training. Approximately 98% of the students in each class participated in the 4-hour examination. The class sizes ranged from 11 to 52. The tests and test sections were presented in a fixed order and were separately timed. Frequent rest intervals were scheduled to reduce fatigue.

The students were advised that the test was for research purposes only, but were encouraged to perform to the best of their ability. After the administration, the students rated their overall effort level. Students who reported an effort level of four or less were eliminated from further analysis. After a preliminary comparison of IEKW scores, students and instructors were informed of the results. The results were then discussed with the students and instructors.

flight hours, and all administrative changes data were collected for each tested student. Finally, archival records were searched to collect RFAST scores for the tested students.

Subjects. Of the 563 tested students in the data base, 532 graduated from IERW, 17 were eliminated for flight or academic deficiencies, and 14 were eliminated for other reasons (e.g., medical, misconduct). Excluding the 14 students eliminated for other reasons, the pre-operational validation data base includes 549 students: 233 are COs and 316 are WOCs; 283 took Form E of the NFAST and 266 took Form F. Complete RFAST scores were obtained for 373 of the tested students.

Most of the CO and WOC students were Second Lieutenants and Sergeants, respectively. The only other substantial difference between the two groups was in the years of education completed. The CO students had completed a median 4 years of college while the WOC students had completed a median of only 1 year of college. A majority of the students were serving in the active Army (63%) with the remainder serving in the Army Reserve (13%) and National Guard (24%).

The median age of the students was 24 with a range of 18 to 36. The majority of the students were male (95.6%). Of the 549 entering students, 12.1% held a fixed-wing license, .4% held a rotary-wing license, and .4% held both types of license. Only a small percentage of the students indicated they disliked participating (5.1%) in the research or preferred some other duty to participating (4.6%) in the research.

Results

The results indicate that most of the NFAST tests assess reliable individual differences in aviation-related abilities. The average difficulty levels are near the optimum of .50, the variances reflect substantial individual differences in ability, and the estimates of reliability are within an acceptable range (see Table 2). The alternate forms of the NFAST are equivalent for five of the tests. The largest difference is between the two forms of the Finding Figures test. Test performance by the CO and WOC students is quite similar, although the CO students scored significantly higher ($p < .01$) on four of the tests.

The Helicopter Knowledge test results indicate that it is not difficult and that there is limited variability in the scores. However, all of the students had previously taken the RFAST version of the test. The WOC students scored significantly higher on the test than the CO students ($p < .01$), but there was no difference in performance by either student group on the two forms of the test. This result indicates that there was no sampling bias in terms of aviation-related knowledge in assigning students to the alternate forms of the NFAST battery.

Table 2

Descriptive Statistics for the NFAST and the RFAST

Test	Items	Form E (n = 283)			Form F (n = 266)		
		Mean	SD	Alpha	Mean	SD	Alpha
Flight Planning	72	34.1	9.03	.84	34.4	8.63	.84
Chart Use*	24	15.5	3.62	.71	16.2	3.51	.74
Figure Orientation	56	32.0	9.56	.94	30.3	10.20	.94
Finding Figures*	80	45.7	15.92	.97	38.2	14.98	.96
Finding Rules	20	11.2	3.37	.70	11.1	3.14	.64
Rapid Match*	64	32.3	5.99	.88	34.1	6.76	.90
Sound Reasoning	20	10.7	2.98	.72	9.9	3.20	.75
Helicopter Knowledge	20	16.1	2.76	.67	16.1	2.64	.63
RFAST (n = 195, 178)	200	117.1	14.24	-	117.8	14.49	-

*Significant difference between Forms E and F ($p < .01$).

A factor analysis of the NFAST sections resulted in an interpretable six-factor solution that accounted for 44.3% of the variance. Except for the Finding Rules and Helicopter Knowledge tests, the sections of each test loaded on one and only one factor. The simple correlations between the predictor tests and the OAG (mean = 88.7, $SD = 2.98$) ranged from .24 to .44 (see Table 3). The Chart Use and Flight Planning tests have the highest relationship to the criterion; the RFAST score has one of the lowest correlations. Only the Chart Use, Flight Planning, Helicopter Knowledge, and Figure Orientation tests entered the stepwise multiple regression equation for the OAG, with $R = .52$, $F(4, 844) = 49.89$, $p < .0001$. Similar results were obtained in separate regression analyses for Forms E and F and with the CG and WU groups.

Conclusions

These results indicate that a subset of tests from the NFAST battery will significantly improve the effectiveness of the selection proceedings used for the USMC course. Using a preselected sample of flight students, 14 of the new tests accounted for 24% of the performance variance in an intense, 6-month training course. The complex of perceptual and perceptual abilities measured by the NFAST battery are important to the most promising instrument for predicting the competencies required in the selection and retention of USMC trainees.

Table 3

Intercorrelation Matrix of Predictor Tests and Overall Average Grade

	FP	CU	FO	FF	FR	RM	SR	HK	RF
CU	54								
FO	38	34							
FF	37	36	43						
FR	31	30	28	26					
RM	34	33	33	36	22				
SR	37	27	22	22	26	23			
HK	14	21	15	19	13	-03	10		
RF	28	25	31	26	25	20	13	40	
OA	41	44	30	29	24	24	24	25	29

Note. FP = Flight Planning; CU = Chart Use; FO = Figure Orientation; FF = Finding Figures; FR = Finding Rules; RM = Rapid Match; SR = Sound Reasoning; HK = Helicopter Knowledge; RF = RFAST; OA = Overall Average. Decimals omitted.

References

- Eastman, R. F., & McMullen, R. L. (1978). Item analysis and revision of the flight aptitude selection tests (ARI Field Unit Research Memorandum 78-4). Fort Rucker, AL: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Kaplan, H. (1965). Prediction of success in Army aviation training (Technical Research Report 1142). Washington, DC: U.S. Army Personnel Research Office.
- Lockwood, R. E., & Shipley, B. D., Jr. (1984). Evaluation of the Revised Flight Aptitude Selection Test (Technical Report ASI479-020-84). Fort Rucker, AL: Anacapa Sciences, Inc.
- McAnulty, D. M., Cross, K. D., & Jones, D. H. (1986). The development of an experimental battery of aviation related ability tests (Technical Report ASI678-201-86[B]). Fort Rucker, AL: Anacapa Sciences, Inc.
- McAnulty, D. M., Jones, D. H., Cohen, R. J., & Lockwood, R. E. (1984). Identification of the abilities required for effective helicopter training performance (Technical Report ASI479-046-84[B]). Fort Rucker, AL: Anacapa Sciences, Inc.

Retaking the U.S. Navy and Marine Corps Aviation Selection Test Battery

Annette G. Baisden
Naval Aerospace Medical Institute
LCDR F. Douglas Holcombe
Naval Safety Center

Abstract

It has been recognized informally that an increasing number of recent training failures, particularly minority students, had taken the aviation selection test several times before qualifying. This study assesses retake performance on the aviation selection test battery, analyzes test performance by race/ethnic groups, and examines retest effects on predicting success in aviation training.

The U.S. Navy and Marine Corps Aviation Selection Test battery is a paper-and-pencil test used as the primary instrument for selecting student naval aviators and student naval flight officers. The test battery consists of four sub-tests: Academic Qualification Test (AQT), Mechanical Comprehension Test (MCT), Spatial Apperception Test (SAT), and Biographical Inventory (BI).

Under current usage, aviation officer selection is based upon two scores: the AQT score and the Flight Aptitude Rating (FAR) score. The AQT is a general aptitude test and is predictive of ground school performance ($r=.29$). The FAR represents a combination of scores on the MCT, SAT, and BI. These tests measure familiarity with mechanical concepts, ability to visualize the relationship between the attitude of an airplane and the territory over which it is flying, personal history, and aviation knowledge. The FAR is predictive of success or failure in the flight training program ($r=.31$).

Policy at the time of the study permits a retest after a six month time interval, with no maximum number of retests specified. About 30 percent of students entering training have taken the test more than once.

Method

Data on all student naval aviators (SNAs) entering aviation training during FY-82 through FY-84 were used for analysis. Most students had either completed undergraduate flight training (wings) or had attrited by early FY-87. Of the 5775 students, 79 (1.4%) are female, 5464 (94.7%) are Caucasian, 123 (2.1%) are black, 74 (1.3%) are Hispanic, 108 (1.8%) are other minorities, and 6 (0.1%) are unidentified.

The students took the aviation selection tests between 1975 and 1984. At the time of testing 54.6 percent were not college graduates; however, all had completed college prior to entering

training. Service status at the time of testing was represented as follows: civilians, (58.9%); Naval Reserve Officer Training Corps, (13.3%); naval academy, (13.9%); other (13.9%).

Race/sex identification was self-reported on the aviation selection test data sheet. Ethnic group data and completion/attrition data were derived from standard flight training data forms.

Results

Table 1 shows the frequency of retakes for each race/ethnic group of student naval aviators. Blacks have the largest percentage of retakes, ranging from 41.9-54.4 percent.

Table 1

Test Frequency by Race/Ethnic Group and by Fiscal Year

		Number of Selection Tests Taken											
Race/ Ethnic	FY	1		2		3		4		≥ 5		Total	
		N	%	N	%	N	%	N	%	N	%	N	
Cauc	82	1561	71.7	503	23.1	87	4.0	24	1.1	4	.1	2179	
	83	1019	67.0	385	25.3	95	6.3	16	1.1	5	.3	1520	
	84	1179	66.9	459	26.0	103	5.8	18	1.0	6	.3	1765	
Blk	82	19	55.9	10	29.4	4	11.8	0	.0	1	2.9	34	
	83	25	58.1	6	14.0	7	16.3	2	4.6	3	7.0	43	
	84	21	45.6	13	28.3	10	21.7	1	2.2	1	2.2	46	
His	82	14	70.0	3	15.0	3	15.0	0	.0	0	.0	20	
	83	17	48.6	13	37.1	4	11.4	1	2.9	0	.0	35	
	84	12	63.1	6	31.6	1	5.3	0	.0	0	.0	19	
Oth	82	17	73.9	5	21.7	1	4.4	0	.0	0	.0	23	
	83	38	63.3	13	21.7	7	11.7	1	1.7	1	1.6	60	
	84	19	76.0	4	16.0	2	8.0	0	.0	0	.0	25	

For Caucasians, there is a slight increase, 4.8 percent, in the percentage of retakes from FY-82 to FY-84. The increase in retakes for blacks is 10.3 percent, with the largest increase occurring for third tests. The number of retests for Hispanics increased 6.9 percent during the same period.

Descriptive statistics showing how test scores change as a function of retakes are provided in Tables 2 and 3. Performance on the AQT and FAR is scaled in stanines. As shown in Table 2, there is a general improvement in AQT scores with retesting. The largest improvement (0.9) occurs for those tested twice. The average score increase on an AQT with three or more tests is about 0.5. The number of retests increases as the score on the original test decreases. Regardless of the number of times tested, the largest score increase occurs with the final test administration.

Table 2

AQT Scores as a Function of Number of Times Tested						
Number of tests	Descriptive Statistics	Test Number				
		1	2	3	4	5
1	Mean	5.42				
	s.d.	1.30				
	N	3912				
2	Mean	4.39	5.21			
	s.d.	1.28	1.23			
	N	1397	1397			
3	Mean	3.70	4.21	4.78		
	s.d.	1.19	1.16	1.13		
	N	320	320	320		
4	Mean	3.56	3.66	4.13	4.63	
	s.d.	1.15	.94	1.11	1.04	
	N	62	62	62	62	
5	Mean	2.75	3.44	3.44	3.94	4.75
	s.d.	1.13	1.21	1.10	1.29	1.48
	N	16	16	16	16	16

Table 3

FAR Scores as a Function of Number of Times Tested						
Number of tests	Descriptive Statistics	Test Number				
		1	2	3	4	5
1	Mean	6.88				
	s.d.	1.55				
	N	3912				
2	Mean	5.00	6.81			
	s.d.	2.17	1.60			
	N	1205	1388			
3	Mean	4.07	5.01	6.56		
	s.d.	2.09	1.94	1.45		
	N	212	272	318		
4	Mean	3.10	4.17	4.50	6.15	
	s.d.	2.15	2.15	2.01	1.33	
	N	31	42	54	60	
5	Mean	4.25	4.00	4.64	4.62	6.69
	s.d.	2.43	1.67	1.86	1.39	1.82
	N	8	11	11	13	16

Compared to AQT retakes, larger increases in test scores occur with retesting on the FAR, as shown in Table 3. As with the AQT, the smaller the initial test score the greater the number of retakes. Again, regardless of the number of times tested, the largest score increase occurs with the final test administration.

Mean score increases or decreases with repeated testing on the AQT and FAR by race/ethnic group are presented in Table 4. These data compare everyone's test scores for the original tests with test scores for the second tests, all test scores for the second test with all test scores for the third tests, etc., without regard to the total number of tests taken by individuals. Initially, FAR scores increase more than AQT scores increase. In general, for each race/ethnic group, both score differences tend to diminish with an increasing number of retakes.

Table 4
Changes in AQT and FAR Scores
with Repeated Testing by Race/Ethnic Group

Race/ Ethnic	Test	Mean Score Increase/Decrease between tests					
		1-2	2-3	3-4	4-5	5-6	6-7
Cauc	AQT	.75(1674)	.56(352)	.38(72)	1.07(15)	-.33(3)	2.00(1)
	FAR	1.75(1321)	1.58(275)	1.44(55)	1.50(12)	1.50(2)	3.00(1)
Blk	AQT	.83(58)	.07(29)	1.50(8)	-.20(5)	-1.00(2)	1.00(1)
	FAR	2.02(45)	1.04(25)	1.17(6)	.00(4)	1.00(1)	.00(1)
His	AQT	.40(30)	.13(8)	2.00(1)			
	FAR	2.04(24)	1.60(5)	.00(1)			
Oth	AQT	.62(34)	.92(12)	.00(2)	.00(1)		
	FAR	1.82(28)	1.56(9)	.50(2)	4.00(1)		

Table 5 shows the completion percentage in Naval Aviation Schools Command (NASC) and primary flight training for SNAs who had taken the selection tests one or more times. As seen in the table, completion percentages for Caucasians are very consistent for students entering training with four or fewer tests. With five or more tests a large reduction in completion percentage occurs. For blacks, percentage of completion is generally less than that of Caucasians and decreases with three or more tests. Hispanics have the highest completion percentage rates of the groups examined with a drop in the rate at the second test.

Discussion

In summary, there is a general improvement in test scores on retests with the larger increases occurring on the FAR. The results suggest that taking the aviation selection test battery up to a total of four test administrations does not lower its ability to predict success in NASC and primary flight training. Percentage of completion decreases substantially with five or

more tests. An analysis by race/ethnic group shows completion rates for blacks begin to decrease after the second test.

Follow-up studies should address issues which could explain increases in test scores. These include motivation for retesting, practice effects, education, maturation, and time intervals between test administrations.

Table 5

Completion Rates by Race/Ethnic Group and Number of Tests							
Stage	Race/ Ethnic	Number of Selection Tests					Total
		1	2	3	4	≥ 5	
NASC	Cauc	78.4 (2948)	79.4 (1069)	80.0 (228)	81.0 (47)	73.3 (11)	78.8 (4303)
	Blk	69.2 (45)	69.0 (20)	47.6 (10)	33.3 (1)	33.3 (1)	62.6 (77)
	His	97.7 (42)	81.8 (18)	100.0 (8)	100.0 (1)	0.0 (0)	93.2 (69)
	Oth	85.1 (63)	81.8 (18)	80.0 (8)	100.0 (1)	100.0 (1)	84.3 (91)
	Total	78.6 (3098)	79.1 (1125)	77.9 (254)	79.4 (50)	61.9 (13)	78.6 (4540)
Pri- mary	Cauc	68.7 (2583)	65.3 (880)	67.0 (191)	67.2 (39)	50.0 (6)	67.7 (3699)
	Blk	53.8 (35)	55.2 (16)	42.9 (9)	33.3 (1)	33.3 (1)	50.4 (62)
	His	76.7 (33)	68.2 (15)	75.0 (6)	100.0 (1)	0.0 (0)	74.3 (55)
	Oth	71.6 (53)	54.5 (12)	60.0 (6)	100.0 (1)	0.0 (0)	66.7 (72)
	Total	68.6 (2704)	64.9 (923)	65.0 (212)	66.7 (42)	33.3 (7)	67.3 (3888)

CROSS-VALIDATION OF AN EXPERIMENTAL PILOT SELECTION AND CLASSIFICATION TEST BATTERY

Thomas R. Carretta, Ph.D.

Aircrew Selection & Classification Function,
Cognitive Skills Assessment Branch, Manpower & Personnel Division,
AF Human Resources Laboratory, Brooks Air Force Base, TX

Since 1955, most U.S. Air Force pilot trainees have been selected on the basis of their physiological fitness, academic performance, aptitude test results, and previous flying experience. Once an applicant has been selected for Undergraduate Pilot Training (UPT), the emphasis shifts from screening to training with the eventual goal of placement in a specialized follow-on training track (fighter or non-fighter aircraft).

The current UPT program lasts 52 weeks and involves a T-37 phase (initial jet trainer, 21 weeks) and a T-38 phase (advanced jet trainer, 31 weeks). During the 43rd week of training, an Advanced Training Recommendation Board (ATRB), which consists of T-38 Instructor Pilots, meets to evaluate UPT students for specialized training assignments. In 1991, the Air Force will replace this program with a Specialized Undergraduate Pilot Training (SUPT) program where the specialized training assignment will be made prior to entry into UPT. Therefore, it will be necessary to classify pilot candidates into specialized training tracks without the benefit of flying training performance measures. Also, once a pilot trainee has entered a particular track, he/she will not be allowed to change to the other track.

Given the expense of pilot training and the different demands associated with fighter and non-fighter aircraft, it is crucial that pilot trainees be assigned to specialized training tracks in an optimal manner.

Preliminary results (Carretta, 1988) from a computerized test battery, known as the Basic Attributes Tests, or BAT, suggest that individual differences in hand-eye coordination, information processing ability, personality and attitudes are related to flying training performance and add to the predictive validity of a currently used selection and classification instrument, the Air Force Officer Qualifying Test (AFOQT). Concerns regarding the stability or the prediction models were raised, because the final selection and classification models were developed using a stepwise regression approach, and there was some evidence of redundancy among the test measures. Therefore, a different final regression solution may have occurred with a different sample. The present study was performed to cross-validate these results to determine the generalizability of the original prediction models.

METHOD

Subjects

The subjects in this study were 709 U.S. Air Force UPT students from the Air Force Reserve Officer Training Corps (AFROTC) and Officer Training School (OTS) who were tested on both the AFOQT and BAT. These subjects already had been chosen for UPT, in part, on the basis of their AFOQT scores.

Instrumentation

AFOQT. The AFOQT is a paper-and-pencil aptitude test battery used to select civilian applicants for officer precommissioning training programs and to classify commissioned officers into aircrew job specialities (pilot vs navigator training). The battery consists of 16 subtests that assess five ability domains: verbal, quantitative, spatial, aircrew interests/ aptitude and perceptual speed (Skinner & Ree, 1987). Fourteen of the 16 AFOQT subtests are used to compute the Pilot and Navigator-Technical composite indices used in the operational selection of pilot trainees (U.S. Air Force, 1983).

Basic Attributes Tests (BAT). The BAT battery consisted of 12 computerized tests that assessed individual differences in psychomotor coordination, information processing, perceptual abilities, personality, and attitudes. The types of scores generated from these tests include tracking error, response time, response accuracy, and speed by accuracy trade-offs. Table 1 provides a brief summary of this battery. A more detailed description is provided by Carretta (1987, 1988).

UPT Performance Criteria. UPT final training outcome was scored as a dichotomous variable with graduates receiving a score of 1 and failures a score of 0. UPT graduates received a recommendation for a specialized follow-on training assignment in either a Tanker-Transport-Bomber (TTB) aircraft or a Fighter-Attack-Reconnaissance (FAR) aircraft. Those recommended for fighter assignments received a score of 1, while those recommended for non-fighter assignments received a score of 0. ATRB recommendation was used as an indicator of training performance, as training grades were not available.

Apparatus

The BAT apparatus consists of a microcomputer and monitor built into a ruggedized chassis with a glare shield and side panels designed to minimize distractions. The subjects responded to the tests by manipulating individually or in combination, a dual-axis joystick on the right side, a single-axis joystick on the left side, and a keypad in the center of the test unit. The keypad included keys labeled 0 to 9, an ENABLE key in the center, and a bottom row with YES and NO keys, and two others for same/left responses (S/L), and different/right responses (D/R).

Procedure

Each subject was administered both the AFOQT and BAT prior to entry into UPT. Pilot trainees were commissioned through either AFROTC or OTS. Those from AFROTC were tested on the AFOQT prior to entering college, or while an undergraduate. AFROTC pilot trainees were administered the BAT while attending a Flight Screening Program (FSP) in the summer following their junior year in college. For the OTS pilot trainees, the AFOQT was administered either near to or after completion of a college degree, and the BAT was administered at the beginning of FSP.

The BAT battery, as used in this study, consisted of 12 tests and required about three and one half hours to complete. After the test administrator initiated the battery, the test session was self-paced by the subjects. Programmed breaks were included between tests in order to reduce mental and physical fatigue.

TABLE 1. BASIC ATTRIBUTES TESTS (BAT) BATTERY SUMMARY

TEST NAME	LENGTH (mins)	ATTRIBUTES MEASURED
Test Battery Introduction:	15	Biographical Information
Two-Hand Coordination: (rotary pursuit)	10	Tracking & Time-Sharing Ability in Pursuit
Complex Coordination: (stick and rudder)	10	Compensatory Tracking Involving Multiple-Axes
Dot Estimation:	6	Impulsiveness/Decisiveness
Digit Memory:	5	Perceptual Speed
Encoding Speed:	20	Verbal Classification
Mental Rotation:	25	Spatial Transformation & Classification
Item Recognition:	20	Short-Term Memory, Storage, Search & Comparison
Risk Taking:	10	Risk Taking
Embedded Figures:	15	Field Dependence/Independence
Time Sharing:	30	Higher Order Tracking Ability, Learning Rate & Time Sharing
Self-Crediting Word Knowledge:	10	Self-Assessment Ability, Self-Confidence
Activities Interest Inventory:	10	Survival Attitudes

All pilot trainees went through the same UPT program which lasted 52 weeks. The advanced training recommendation was made in the 43rd week of training, and the final training outcome was determined at the end of the program.

Approach

To be useful as an adjunct to currently-used pilot selection and classification methodologies, the BAT performance measures must demonstrate a significant increment in predicting training outcome when used in addition to operational instruments (i.e. AFOQT scores). To evaluate the incremental validity of the BAT, 709 pilot trainees were divided randomly into two groups. The assignments were made so that the groups were similar in their UPT pass/fail rate and fighter/non-fighter recommendation rate. Pilot selection and classification models were developed independently for each group using UPT pass/fail and fighter/non-fighter recommendation for selection and classification criteria, respectively.

In each group, three regression models were evaluated against UPT final outcome and ATRB recommendation. The first model included only AFOQT Pilot and Navigator-Technical composite percentile scores. It served as a baseline by which to judge the incremental validity of the BAT performance measures. The second model used an approach that forced the two AFOQT composites to enter the regression equation first, and then allowed the BAT scores to enter in a stepwise manner. The stepwise results were examined to identify predictor variables common to both groups and form a final model to be used in the cross-validation phase. Regression weights from each sample were applied to the other sample to cross-validate the models.

RESULTS

Prediction of UPT Final Outcome

Table 2 presents results of regressing various predictor combinations on UPT outcome. The objective was to identify the best combination of predictors to use in support of selection decisions.

As shown in Table 2, the AFOQT scores were not related strongly to UPT final outcome for either Group 1 ($R=.126$, $p \leq .10$), or Group 2 ($R=.155$, $p \leq .05$). The magnitude of these relationships may be due, in part, to a restriction in range on the aptitude domain, measured by the AFOQT, as these subjects had already been screened for UPT on the basis of their AFOQT performance.

Results from the stepwise regression analyses (not shown in Table 2) suggested that individual differences in performance on the BAT battery were related strongly to final training outcome (Group 1: $R=.452$, $p \leq .01$; Group 2: $R=.378$, $p \leq .01$). A comparison between the stepwise UPT models for the two groups indicated that they shared 12 common predictors from the AFOQT and six BAT tests. These included the AFOQT Pilot and Navigator-Technical composites, three tracking error scores from the psychomotor tests, two cognitive test scores (Item Recognition and Encoding Speed response times), and five personality/attitudinal test scores (Self-Crediting Word Knowledge and Activities Interests Inventory). A reduced model that used only these 12 scores was related significantly to UPT final outcome for both groups (Group 1: $R=.303$, $p \leq .01$; Group 2: $R=.342$, $p \leq .01$). Subjects with good hand-eye coordination who made quick decisions and exhibited a cautious risk-taking strategy were more likely to complete training successfully. When the regression weights from each sample were applied to the other sample to cross-validate the models, some reduction in the validity coefficients was observed. However, the cross-validated models were statistically significant (Group 1: $r=.176$, $p \leq .01$; Group 2: $r=.220$, $p \leq .01$).

A final set of analyses was performed using the entire sample (Group 1 and Group 2 combined). The 12 variable AFOQT/BAT model ($R=.290$, $p \leq .01$) was related more closely to UPT final outcome than a model that used only the AFOQT ($R=.150$, $p \leq .01$) ($F [10,696] = 4.68$, $p \leq .01$).

Prediction of Fighter/Non-Fighter Recommendation

Table 3 summarizes the cross-validation analyses for the fighter/non-fighter recommendation. AFOQT performance was not related strongly to follow-on training recommendation for either group; (Group 1: $R=.197$, $p \leq .01$; Group 2: $R=.059$, n.s.).

The stepwise AFOQT/BAT regression approach (not shown in Table 3) yielded impressive multiple correlations with follow-on training recommendations in both groups (Group 1: $R=.592$, $p \leq .01$; Group 2: $R=.565$, $p \leq .01$). However, it appeared that the final regression solutions over-fitted the data. That is, many of the variables added little to the prediction of follow-on training recommendation.

A reduced model was evaluated that used the two AFOQT scores and six scores from four BAT tests that appeared to contribute to the prediction of performance in both groups. These included three learning rate and dual-task performance scores from the Time Sharing test, and one score each from three cognitive abilities tests (Item Recognition, Encoding Speed and Mental Rotation response times). This model yielded mixed results with the two

groups. The multiple correlation was strong in Group 1 ($R=.336$, $p \leq .01$), but only marginally significant in Group 2 ($R=.202$, $p \leq .10$). However, when the regression weights from each group were cross-validated with the other group, only minor reductions in the validity coefficients occurred, and the correlations were significant for both groups (Group 1: $r=.302$, $p \leq .01$; Group 2: $r=.188$, $p \leq .01$). When applied to the full sample, the reduced AFOQT/BAT model was significant ($R=.267$, $p \leq .01$) and added to the predictive utility of the AFOQT ($R=.126$, $p \leq .05$) ($F [6,484] = 4.81$, $p \leq .01$). Subjects with quicker learning rates and response times were more likely to be recommended for follow-on training with fighter aircraft.

DISCUSSION

Results from the AFOQT and BAT cross-validation regression analyses were encouraging. For each group, individual differences in psychomotor skills, cognitive and perceptual abilities, and personality and attitudinal characteristics helped to reduce uncertainty in making pilot selection and early training classification decisions. Further, results from the model development and cross-validation phases suggest that the selection and classification models are robust. The significance of the cross-validation analyses is especially important, as it provides an estimate of the expected validity of the selection and classification models in an operational setting.

The validity estimates provided by the cross-validated selection models may seem low (Group 1: $r=.188$, $p \leq .01$; Group 2: $r=.220$, $p \leq .01$). It should be noted, however, that the magnitude of these correlations was limited by the dichotomous nature of the UPT outcome measure (pass/fail) and by the proportion of UPT graduates in the sample (66.7 % graduates). A more sensitive outcome measure (i.e. class ranking) may yield a larger validity coefficient.

Validity estimates provided by the cross-validated classification models suggest that the advanced training recommendations made by the T-38 Instructor Pilots during the 43rd week of UPT can be predicted prior to entry into UPT with reasonable success. Additional research needs to be conducted, however, to develop more suitable criteria for the development of classification models. Current plans are to explore the availability and suitability of training performance data in specialized training tracks for this purpose.

SUMMARY

In this study, a combination of AFOQT and BAT performance measures demonstrated utility for improving USAF pilot selection and reducing uncertainty when making recommendations for specialized training assignments.

The model development phase indicated substantial agreement between the selection and classification models that were developed independently for the two groups. Further, the cross-validation phase demonstrated that these models are robust enough to be used as adjuncts to currently-used USAF pilot selection and classification methodologies.

TABLE 2. PREDICTION OF UPT OUTCOME

SAMPLE/MODEL	N	N SCORES	UPT PASS RATE	R	r ^a
GROUP 1:					
AFOQT only	333	2	.682	.126 n.s.	
AFOQT and BAT (reduced)	347	12	.677	.303++	.176++
GROUP 2:					
AFOQT only	345	2	.655	.155+	
AFOQT and BAT (reduced)	362	12	.658	.342++	.220++

TABLE 3. PREDICTION OF ATRB OUTCOME

SAMPLE/MODEL	N	N SCORES	FAR RATE	R	r ^a
GROUP 1:					
AFOQT only	247	2	.544	.197++	
AFOQT and BAT (reduced)	254	8	.563	.336++	.312++
GROUP 2:					
AFOQT only	228	2	.595	.059 n.s.	
AFOQT and BAT (reduced)	239	8	.565	.202 n.s.	.188++

Note: Sample sizes within Group 1 and Group 2 vary, because some subjects did not complete all tests from the BAT due to periodic changes in the battery.

^aThe column labeled "R" indicates the multiple correlation of the model based on the regression weights from that group. The column labeled "r" indicates the correlation of the predicted outcome with actual outcome based on the regression weights from the other group (cross-validation).

n.s. (non-significant) | +p ≤ .05 | ++p ≤ .01

REFERENCES

- Carretta, T.R. (1987). Basic Attributes Tests (BAT) System: Development of an Automated Test Battery for Pilot Selection, AFHRL-TR-87-9, Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Carretta, T.R. (1988, in press). USAF Pilot Selection and Classification Systems. Aviation, Space, and Environmental Medicine.
- Skinner, J. & Ree, M.J. (1987). Air Force Officer Qualifying Test (AFOQT): Item and Factor Analysis of Form O, AFHRL-TR-86-68, Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- United States Air Force (1983). Application Procedure for IPT, UPTH and UNT, Air Force Regulation 51-4, Washington DC, Department of the Air Force.

Current Directions in Air Force
Pilot Selection and Classification Research¹

Frederick M. Siem

Air Force Human Resources Laboratory
Brooks AFB, Texas

Dwight C. Hageman

Metrica, Inc.
San Antonio, Texas

Theresa A. Mercatante

Air Force Human Resources Laboratory
Brooks AFB, Texas

I. INTRODUCTION

Employment interviews are widely used in industry and the military for personnel selection and classification, although the scientific evidence for their validity and reliability is somewhat weak (Arvey & Campion, 1982; Harkness, 1987). The U.S. Air Force does not currently use a routine interview process for pilot selection, other than that conducted for recruiting purposes, although some components of the Air Force do rely on their own procedures. Units of the Air National Guard, for instance, have used an interview process with some success (Armour, personal communication, October 19, 1987).

Based on the perceived success of an interview process used by the Air National Guard, a study was undertaken to assess the validity of an interview for selecting Air Force pilot candidates. In particular, the issue of interest was whether instructor pilots could use a structured interview technique to assess an applicant's potential for pilot training. The content of the interview was derived from a review of the instrument used by the Air National Guard and through discussions with subject matter experts of the Air Force Air Training Command (ATC), the organization responsible for training all Air Force, Air National Guard and Air Force Reserve pilot candidates.

1. The opinions expressed herein are those of the authors and do not necessarily reflect those of the United States Air Force. The authors wish to acknowledge the assistance of Dr. James E. Campion, University of Houston, and of Martin J. Dittmar, Metrica, Inc., in the design of the interview instrument and in conducting interviewer training.

II. METHOD

Respondents

The sample consisted of 112 pilot candidates participating in a Flight Screening Program (FSP) conducted with cadets from the Air Force Reserve Officer Training Corps (AFROTC). Most of the sample were between their junior and senior year in college, and all but two of the cadets were male.

The interviewers were ten Air Force Instructor Pilots who had received about eight hours of instruction in interview techniques. The instruction consisted of explanation of the purpose of each question, followed by group discussion of both actual and hypothetical responses to each question so as to establish standardized rating guidelines. Instructor pilot experience with training undergraduate pilots ranged from 30 to 102 months, and their time flying jets varied from 2500 to 3600 hours.

Instruments

The interview consisted of 40 questions designed to collect three types of information. Seven background questions provided information concerning work and academic history. The bulk of the interview consisted of 26 questions related to motivation and self-confidence factors, such as involvement in aviation-related activities, evidence of goal-setting behavior, and participation in competitive group activities. The final seven questions were designed to determine the extent to which the cadet was familiar with the job of being an Air Force pilot. On the basis of responses to the questions, the instructor pilot used a seven-point scale to rate the cadet on five aspects: motivation, self-confidence, job-related knowledge, potential to complete training, and potential to perform well in training.

Procedure

Prior to the beginning of FSP, one instructor pilot interviewed each cadet in the sample. The process took from thirty to forty-five minutes to complete. The interviewer read each question exactly as it appeared in the interview form. If the cadet did not understand or failed to respond, the interviewer repeated the question a second time. Interviewers were instructed to respond to 'Yes' or 'No' answers by asking the cadet to elaborate, in order to give the interviewer enough information to make a rating.

Subsequent to the interview, cadets participated in the FSP that included 14 hours of flight time in the military version of a Cessna 172. The screening program lasted about three weeks, at which time two criterion measures were collected. One criterion was whether the cadet graduated from the program. Cadets could fail to complete the training program for a number of reasons, including poor academic or flying performance, medical problems

or air-sickness. Some cadets also determined after the first few flights that they were not interested in pursuing further the rigors of pilot training and left the program voluntarily.

The second criterion was a rating assigned by an FSP representative on a five point scale based on a review of each cadet's academic and flight performance records. Non-graduates were assigned the lowest score (one). Graduates were rated from two to five, with two being poor performance and five representing outstanding performance.

III. RESULTS

Descriptive statistics for the variables included in this study are provided in Table 1. Note that the training completion ratio in this sample was .74. Of the 29 non-graduates, the majority (17) were eliminated due to flying training deficiency (FTD). Only one cadet failed to complete training due to manifestations of apprehension (MOA), whereas eight were self-initiated eliminees (SIE) and three failed to graduate due to medical or academic problems.

Table 1.

Descriptive Statistics for Interview Rating Items and Training Outcomes ($N = 112$)

Measure	Mean	S.D.	Range
Interview Rating Items			
Motivation	4.63	1.36	1 - 7
Self-confidence	4.96	1.40	1 - 7
Knowledge	4.08	1.47	1 - 7
Potential to complete training	4.71	1.37	1 - 7
Potential to perform well in training	4.49	1.37	1 - 7
Training Outcomes			
Training Performance	2.66	1.25	1 - 5
Training Completion	.74	.44	0 - 1

Note. Higher numbers indicate more desirable characteristics and performance outcomes.

Validity coefficients for five of the interview rating items are displayed in Table 2. The coefficients represent correlations with two criterion measures: a five-point performance rating (with five representing superior performance and one representing non-graduation), and a dichotomous measure of training completion (pass/fail). The predictor variables consisted of interviewer summary ratings of cadet motivation, self-confidence, job-related knowledge, potential for completing flight training, and potential to perform well in training.

Table 2. Correlations Between Interview Ratings and Flight Training Performance Criteria (N = 112)

Rating Measure	Correlation	
	Training Performance	Training Completion
Motivation	.13*	.07
Self-confidence	.19*	.08
Knowledge	.18*	.07
Potential to complete training	.20*	.10
Potential to perform well in training	.22*	.15

Note. Results for training completion are point-biserial correlations.

* $p < .05$ that results are due to chance differences from zero

The results in Table 2 indicate that the correlations of the ratings with the five-point performance rating were uniformly higher than those with the dichotomous pass/fail measure. The pattern of results for both criterion measures was similar. However, none of the point-biserial correlations for the pass/fail measure were significantly different from zero, whereas all but one of the rating measures were significantly correlated with the five-point performance measure.

Finally, a factor analysis was conducted to explore the underlying factor structure of the interview information. The results of a principal components analysis suggested that all the variables were measuring a single 'interviewer impression' dimension, as only one factor emerged with an eigenvalue greater than one, explaining nearly 78 percent of the variance.

Examination of an alternative two factor solution produced weak evidence for a second factor explaining 11 percent of the variance (Table 3).

Table 3. Factor Analysis of Interviewer Ratings (N = 112)

----- Factor Matrix -----		
	I	II
Percent of variance explained	77.68	11.10
Measure		
Motivation	.90	
Self-confidence	.89	
Knowledge	.72	.69
Potential to complete training	.90	
Potential to perform well in training	.88	

Note. Only loadings > .20; shown: Factors orthogonal, signs positive

DISCUSSION

The results of this study suggest that instructor pilot impressions, formed during an interview using a structured format, appear to be valid predictors of FSP performance. The observation that the summary ratings are accounted for by a single factor suggests that the interview may be measuring a single construct, although such a conclusion awaits more thorough psychometric analyses of ratings data.

Analyses currently underway with the data from this study will address several other issues. Data for inter-rater reliability analysis have been collected and are being analyzed to examine whether the instructor pilot interviewers all used the same psychological criteria in rating cadet responses. Other analyses will be conducted to examine the utility of composite measures derived separately for both the set of predictor variables and from the set of criterion measures. The results from these analyses are expected to improve substantially proposed aircrew selection and classification systems.

REFERENCES

Arvey, R. D., & Campion, J. E. (1982) The employment interview: A summary and review of recent research. Personnel Psychology, 35, 281-322.

Harkness, S. R. (1987, June). Use of the interview in military selection: A review of one-to-one interview usage as a selection device in TTCP nations. Unpublished report, Technical Panel 3 (UPT-3), The Technical Cooperation Program.

ISSUES IN JOB SAMPLE TESTING

Herbert George Baker, PhD
Gerald J. Laabs, PhD
Navy Personnel Research and Development Center

ABSTRACT

At first glance, developing and administering a test that includes actual job tasks would seem to be a fairly straightforward enterprise. However, it is neither simple nor easy. Development, administration, and scores interpretation of job sample tests in connection with the Joint-Service Job Performance Measurement and Enlistment Standards Project have exposed many technology gaps. Resolution of these problem areas promises to be one of the most significant contributions of military psychology to the theoretical structure of the discipline and to the practical aspects of human performance measurement. This paper delineates a number of technical and practical issues, and discusses progress toward their resolution.

INTRODUCTION

Measurement of that aspect of human performance known as technical proficiency, through the use of hands-on (or job sample) tests, is at the core of the Joint-Service Job Performance Measurement/Enlistment Standards Project. As part of its contribution the Navy is developing job sample tests for eight enlisted jobs.

Because it presumably has the greatest fidelity with respect to the actual content of the job, the hands-on test has been established as the benchmark against which other types of tests and measures will be evaluated. This fact heightens the importance of hands-on testing to the Navy's JPM Program and increases the degree of scrutiny to which both developmental procedures and eventual results will be subjected.

It may seem that developing, administering, and interpreting the results of tests that incorporate actual job tasks poses a rather clear and unambiguous problem. In fact, one might think it easier in some ways than traditional testing, in that genuine job content is used instead of its verbal or photographic representation, with a subject's responding by doing something as opposed to verbalizing, identifying representations, or indicating a choice among answer alternatives. Real tasks are performed on actual equipment in a work space, and success or failure on those tasks should be obvious.

However, job sample testing proves to be neither simple nor easy. As can be imagined by anyone familiar with the operational military environment, the challenges of such things as obtaining a sufficient sample of test subjects, access to and use of expensive and easily damaged equipment, security and readiness considerations, as well as the heavy impact of the testing on military units can be formidable. In addition to the logistical and administrative problems inherent in job sample testing, the development, administration, and scores interpretation of such tests in the Joint-Service Project have also exposed many technology gaps. It is the issues presenting themselves in this area (i.e., technical issues), that we would like to address.

No comprehensive technology exists for the development of hands-on performance tests that is comparable to that available for more traditional measures, such as pencil-and-paper, multiple choice tests. Standards are only beginning to evolve in the JPM arena. Comprehensive and universally accepted standards in the form of a "cookbook" will not emerge soon, although the Joint-Service Project will most certainly contribute to that end.

The opinions expressed herein are those of the authors, are not official, and do not represent Department of the Navy policy.

REPRESENTATIVE ISSUES

Technical issues arise in five major areas. These are: (1) domain specification, (2) test development, (3) test administration, (4) relating results to predictors, and (5) the question of surrogates.

DOMAIN SPECIFICATION

Problems present themselves at the very outset. No person can be tested on all aspects of any job. Even if all of those aspects could be specified -- and they cannot be -- the measurement of individual performance on all tasks included within a job would be temporally and fiscally infeasible. Thus, the initial challenge is to extract the essence of the job. The expense of hands-on test development as well as the often extreme length of the resulting performance tests only intensify this requirement for careful domain specification, which ultimately resolves into the problem of task selection.

Some sample of tasks must be chosen that will faithfully represent the major elements of the job. The set of tasks must also form a comprehensive enough sample such that knowledge of a person's performance on the sample of tasks can be generalized to performance on the entire set of tasks that comprise the job. This process has been denoted as that of narrowing the job content universe, through successive steps, to the test content domain.

The effort often begins by deleting tasks which pose inherent dangers to personnel, cause excessive wear or damage to equipment, violate national security measures, or have an adverse impact on readiness. Additional steps are those of deleting tasks that require joint action by two or more persons and those that do not tap technical proficiency (e.g., workspace cleanup, watchstanding, etc.).

But the difficulties have only begun when the above mentioned tasks have been removed from consideration. Unfortunately, there are no generally accepted methods for gathering information from SMEs and/or determining "critical" tasks. Traditional task selection has relied heavily on the judgment of subject matter experts (SME). We did this in our Navy radiomen performance testing with repeated SME workshops and a multiple hurdles approach, plus a quality control review panel (QCRP) of senior job experts, attempting to ensure test reliability.

In contrast, the National Academy of Sciences' (NAS) Committee on the Performance of Military Personnel has recommended random sampling, so that results can be generalized to the entire job content. This interposes an additional thorny problem into the already difficult job of task selection. However, the Services disagree with the NAS recommendations; in fact, the committee itself is split on the issue.

In our own work, random sampling has been undertaken, but with resultant task samples that are often useless and usually unacceptable to SMEs. The answer to this question of purposive versus random sampling task selection will no doubt be some marriage of convenience, perhaps in the form of stratified random sampling, a process whereby expert judgment defines certain parameters within which randomization of task selection occurs. We have an exploratory effort in progress.

TEST DEVELOPMENT

Here, again, there is a technology deficiency. Traditional test construction and evaluation methodologies exist, but there is a need to more directly address hands-on phenomena. Just about all of the usual considerations of test development remain important, while several others are added to the list.

Incidentally, the process mentioned above, that of narrowing the job content universe to a test content domain, exerts an impact here. To the degree that testing conditions impose themselves on job sample test development and administration, there is a corresponding decrement in fidelity vis a vis the real job world. These can be minimized, but non-job elements (e.g., scoring procedures and rules for generating test stimuli and responses) are inevitable intruders.

However, beyond these, many practical considerations come into play in job sample testing, especially when using real equipment, certain considerations precluding inclusion of particular tasks (or portions of tasks), or forcing an unusual order onto task element performance. These include part-task versus whole task considerations, task sequencing, safety precautions and operational constraints.

Validity of hands-on tests must be approached from several angles. First, there is what is often termed face validity. Regardless of its importance or lack thereof to the world of psychology, it is face validity on which our job sample tests will stand or fall in the eyes of managers and SMEs in the military world.

More significantly, at least from a scientific standpoint, is content validity. This requires that the performance measure address the knowledges, skills, and abilities required by the job, in our case, required by classification standards. The validity of hands-on tests ultimately rests on the twin foundations of job/task analysis and critical task selection. Still further, a crucial dimension of the validity issue is construct validity, meaning the accuracy of the performance measure. In the end, it may prove to be the most critical thing against which JPM efforts are judged.

Reliability issues in performance measurement cannot be resolved easily; again, the technology is lacking. For example, test/retest is not useful, because parallel forms of items addressing task performance usually do not exist; there is often only one way to do something, and any artificial deviation for the purpose of a parallel item results in distancing from the customary job realities, making equating highly problematic. Moreover, the act of completing a hands-on test provides immediate feedback on correct and incorrect procedures, a learning trial that will necessarily affect any retest.

Furthermore, measures of internal consistency are of unknown utility as regards job sample tests, which, because of job design are necessarily heterogeneous. Just what does an alpha coefficient of whatever magnitude tell us in the case of job sample testing?

And, to cite another source of concern, scoring problems are not always easily resolved. To be sure, if performance can be carefully observed and if the observer can really know all that a test subject is doing while performing a task, procedural tasks may be scored step by step, in a straightforward go/no go fashion. Alas, the same is not true of such things as troubleshooting tasks. Here, there may not be any single place at which to begin, or no one way to solve the problem, or any reasonable length of time for task accomplishment. This brings interesting -- one should say knotty -- scoring problems that affect everything from test design to results interpretation.

Face validity, and to some extent content validity, we have addressed through our careful, consistent use of SME workshops and QCRPs. Our aim has been dual: to ensure favorable reception of the performance measures in the field, and to sample to the maximum extent possible the range of tasks (and therefore knowledges, skills, and abilities required) in the Navy jobs for which we are designing job sample tests. Construct validity issues remain largely unresolved except through comparison among multimethod approaches (discussed below).

We have been particularly concerned with the kinds of reliability that we can measure, such as inter-scorer reliability which we assessed on several occasions. To do this meant that we had to have two or more scorers, but we felt it was worth the expense, and our inter-scorer reliabilities are excellent.

TEST ADMINISTRATION

Having, by whatever means, devised a set of measures of performance, the next step is to administer it. Training of hands-on test observers and standardizing the field-test conditions are major concerns. Diligence in these two areas is necessary if extraneous variance in test results is to be minimized.

Question: Who will administer the test? The use of complicated, even esoteric, equipment frequently requires that those who are to administer the test be experts in the equipment on which it will be administered. They alone, sometimes, can tell whether the task is being performed properly; determine whether the results are correct, and be alert to imminent dangers that escape a researcher's notice.

Obviously, this means using SMEs as test administrators. However, SMEs are not research-oriented; they are, rather, training-oriented, having a natural bent to express approval or disapproval, or to coach -- all things which cause havoc in testing! Cost considerations also militate against the employment of civilian test administrators, while the use of military SMEs, in addition to increasing personnel requirements and further weighting the test administrator training burden, may interpose questions of possible familiarity with test subjects, concern with command "showing," etc.

This necessitates very thorough test administrator and observer training, rigid observation of test administration, and, frequently, refresher training or other interventions. There has to emerge a smooth working team consisting of job experts and researchers.

Standardization of testing conditions is also not easy. This is particularly true in the Navy. Even on shore, every site differs somewhat in equipment model, layout, or both. Each Navy ship, of course, is somewhat unique; even the same class of ship built at the same shipyard may include variations in equipment or its positioning, or lighting, or accessibility for either task performance or the observation of that task performance. And, some measures just cannot be administered on real equipment, or in actual working spaces.

To counter these problems, we have had to carefully select our test administrators, conduct very thorough and repeated test administrator training, and continually monitor test administration with an alert eye to things which might introduce extraneous variance. We have also had to construct our tests in such a way that minor differences in equipment configuration or layout do not differentially affect test performance. If a task had to be performed external to the work setting, we have tried to make it as realistic and true to regular procedures as possible. Initial tryouts and more extensive pilot testing of the performance measures have been found necessary.

RELATING RESULTS TO PREDICTORS

In the task of relating performance on the job sample tests to predictors, traditional rules of thumb may prove useful. However, seldom do we find predictor bases that were designed to predict on-the-job performance; rather, they typically were devised to predict tenure, offense-free service, and/or successful completion of an entry-level training program. This fact only adds to the problem of interpreting the results of hands-on tests; in the case of ASVAB validation, it must cause some of the questions of validity to remain moot, pending augmentation of the predictor base.

Another question pertains to the level at which hands-on test results must correlate with predictors. That is, is it the total score alone which is important, or must the results on each task also be in a positive relationship with the predictors. What about the case where the total score is favorably correlated but one or more task (item) results are not?

The Navy has followed the NAS recommendations regarding the analyses necessary to determine the relationship between predictors used in military selection and classification and performance test results. Nevertheless, the degree to which hands-on test results must correlate with predictors is, in the end, a policy question. And the total score versus item score relationships remain unresolved.

THE QUESTION OF SURROGATES

Surrogates (or substitutes) for job sample testing have more than a peripheral interest. Each Service is developing potential surrogates, and, on its part, the Navy has developed several: rating scales, written job knowledge tests, paper-and-pencil simulation tests, videotape-based tailored response tests, and interactive videodisc measures. Obviously, there are several things driving the efforts to produce surrogates, chief among them reduced cost and reduced administration difficulties of surrogates vice job samples.

Results are mixed. It is not our purpose here to discuss the relative merits of different surrogates. However, we might mention that, while surrogates might assist us somewhat in addressing our job sample test construct validity questions, they pose new problems.

For example, what must be the correlation between the potential surrogate and hands-on tests if it is to become regarded as an acceptable substitute? Does only the total score correlation matter, or must the item-level correlation also be strong and positive? And, in the case of the military, must the surrogate correlate not only with its job sample test referent, but also with the predictors? Also, can there be a truly adequate surrogate for a job sample test? One Navy study showed evidence of unique hands-on variance.

Finally, because they must be evaluated against job sample tests, potential surrogates have to be administered to the same persons. However, both tests necessarily provide practice for each other, and when administered together inescapably have an impact on task performance on the other measure. In the Navy project, we have taken pains to counterbalance test administration. In addition, we have measured order of administration effect in several of our job performance studies, and remain alert for practice effects.

CONCLUSION

Resolution of these problems, and others, while it will not be easy, promises to be one of the most significant and lasting contributions of military psychology to the theoretical structure of the discipline and to the practical aspects of human performance measurement. We make no pretensions that these or a myriad of other technical issues have been solved, or are near solution.

Nevertheless, the contributions of the Joint-Service Project are not inconsiderable. The discipline has long been in need of a concerted, wide ranging effort, with sufficient sample sizes. Results are coming in, and we look to that time in the not too-distant future when there will be a coalescence of hands-on performance measurement technology.

**A SYSTEMS APPROACH
TO
THE SYNTHESIS OF MEASURES OF EFFECTIVENESS**

Edward Connelly

Communications Technology Applications, Inc.
7927 Jones Branch Drive, Suite 300
McLean, Virginia 22102

Abstract

A weapon system Measure of Effectiveness (MOE) must indicate the weapon system's contribution to the effectiveness of its unit performing an assigned mission. To insure relevant assessment of the weapon system, an MOE for the unit performing a mission must be established first; then based on that unit MOE, the contribution of the weapon system to the unit effectiveness determined. To assess the weapon system effectiveness in isolation from its unit can result in system assessment criteria not totally relevant to the unit effectiveness.

A unit MOE, in turn, must produce a mission effectiveness score that is consistent with the effectiveness preference of the commander. The commander considers certain factors and trade-offs among those factors when assessing the unit's effectiveness on a mission. If the MOE used to assess unit effectiveness on training or actual missions does not provide effectiveness scores that are consistent with those of the commander, the MOE cannot be accepted as a valid measure.

This paper describes a method of synthesizing a unit MOE and applying the unit MOE to assess a weapon system effectiveness. In addition the method insures that unit effectiveness scores are consistent with those of the commander. The method is illustrated, in this paper, by an application to a tank reconnaissance (movement to contact) mission. Two types of mission success statements are defined. 1) success defined by an event and 2) success defined by the degree to which a specified condition is achieved. The paper also describes a procedure for determining: the factors included in the unit MOE, the variables used to measure each factor, and the critical man/machine/environment parameters that impact on the measure variables. Finally, the paper outlines a method for establishing a minimum acceptable threshold of unit effectiveness and a means for translating that threshold into minimum acceptable effectiveness thresholds for each system comprising the unit. If each weapon system meets or exceeds its assigned effectiveness threshold, the unit is guaranteed to meet or exceed the minimum acceptable unit effectiveness threshold.

A PERFORMANCE MEASUREMENT PROBLEM

Suppose your supervisor says to you: "I completed the effectiveness appraisals for all members of the group by considering the effectiveness of each member individually. Based on that appraisal, you were rated first. But,

I also selected an MOE (a mathematical function of all the factors I consider important when assessing effectiveness of each individual group member) and with the measure you were rated last"! Obviously, you must point out that there is an serious deficiency in the effectiveness appraisal measure your supervisor selected.

But, note that if the supervisor did not evaluate and rank order each member of the group according to performance, there would be no "test" of MOE the supervisor selected. Without the test, MOE might be (unknowingly) used to incorrectly assess performance of group members leading to promotions, salary increases, bonuses, etc. to the wrong group members.

In general, if the mathematical function used for the MOE is to be considered "correct" it must assign scores to each performance demonstration that are not considered different (by some comparison test) from the scores assigned by an individual (or individuals) accepted as the authority in the subject matter. When an authority assigns effectiveness preference scores to each of a set of performance demonstrations, that scored set constitutes a specification of the discrimination requirement of an acceptable performance measure. Any proposed or selected MOE can be tested by determining if the scores it assigns to the performance demonstrations are not different than those of the authority. And further that test can be used to synthesize an MOE that does assign scores not different from those of the authority.

This test of an MOE is typically never performed -- perhaps because investigators feel that their performance measure is "obviously correct" and therefore it needs no test. The question is: what percentage of selected measures will fail such a test? The answer is: probability most of them! The only condition where an MOE that will likely pass the test can be selected is a condition that never exists -- where the effectiveness is measured by a single variable, sampled only once during the exercise and the exercise is performed only once. But, it is the only condition where there are no trade offs among variables, or among multiple samples of a given variable, and since the exercise is performed only once, there are no sequence (learning or fatigue) effects to consider. If the reader has some doubt about the need to test measures - even those that seem to be "obviously correct", we give a reference (Connelly, 1983) that illustrates that two favorite measures "Root-Mean-Square (RMS)" as used in aircraft flight control and "Time-on-target" as used in weapons pointing tasks which seem to be acceptable measures actually fail the test described above.

METHOD OF SYNTHESIZING MOES THAT DO PASS THE TEST

In the following, a method for synthesizing MOEs that do pass the test described above is presented in the context of synthesizing an MOE for a military unit performing a specified mission. First, let's introduce some terms: A Unit MOE Score for a specified mission, reflecting the Quality of Unit Performance is calculated from a Quality of Performance Rule which is a mathematical function of Critical Factors/Tasks which in turn are measured with Factor/Task Rules which are Mathematical Functions of Measurement Variables. A System which is a component of the unit provides performance characterized by critical MAN/MACHINE/ENVIRONMENT PARAMETERS (CMME). Measurement Variables are influenced by (system related) critical man/machine/env. parameters (CMME).

Some examples are taken from an example armored mission -- a single tank performing a reconnaissance (movement to contact) mission.

MOE Score for Mission = 90.

Critical Factors/Tasks: Select route, Concealment during movement, etc.

Quality of Performance Rule = -110 + 2.00 route selection score
+ 1.00 cover/concealment score + etc.

Measurement Variables (for select proper route): Potential speed of movement along route, Percent of route concealed from enemy's direct view,

Factor/Task Rule (For select proper route) = -40.8 + .075 Potential speed of movement + .183 % of route concealed + etc.

CMME Parameters

Environment: Slope, Soil hardness, Obstacle size ...

Equipment: Max speed of equipment, Suspension ...

Man: Max G load tolerance, Vibration tolerance ...

OUTLINE OF THEORY

1. Assessment of the performance of a system, which is a component of a unit, must be based on the impact of system performance on unit effectiveness.
2. To assess system performance, it is first necessary to produce a Quality of Performance Score Rule (QPSR) for assessing the unit's effectiveness in accomplishing a given mission and then to assess the impact of system performance on unit effectiveness using the QPSR.
3. The ultimate source of information on unit QPSR is the unit authority's (SME's) quality of performance preference.
4. The authority (SMEs) can reliably assess quality of each individual performance demonstration (exercise) as it is observed or described.
5. The authority (SMEs) does not know completely and accurately the rule the authority uses to assess the quality of each individual performance demonstration.
6. The Quality of Performance Score Rule (QPSR) the authority uses to assess unit quality of performance is implied by the scores the authority assigns to individual performance demonstrations.
7. The QPSR can be synthesized from the quality of performance scores assigned by the authority to individual performance demonstrations.
8. The authority should obtain and consider independent assessments of unit quality of performance from other SMEs such as the authority's staff.

M E T H O D
ESTABLISHING AUTHORITY'S QUALITY OF
PERFORMANCE (QP) PREFERENCE

Note: the data given as examples in the following steps are taken from a synthesis of a MOE for a single tank on a reconnaissance mission using a tank commander as the authority.

A, STEPS TO SYNTHESIZE MATHEMATICAL FUNCTIONS FOR SCORING EACH FACTOR/TASK

1, Authority identifies critical factor/tasks for the mission.

Example of a critical factor/task for reconnaissance mission: "Select Proper Route"

FOR EACH FACTOR/TASK:

2, Authority identifies measure variables used to measure the quality of performance of that factor/task. Examples of measure variables for factor/task "select proper route":

- a, "Potential Speed of Movement"
- b, "% of route concealed from enemy"

3, Authority identifies a specific unit including a man/machine/env. for the mission -an existing unit that the authority has first hand experience. That unit is called the baseline unit. The authority then identifies values for each measure variable for the baseline unit, and also identifies the largest and smallest values expected for each measure variable:

Examples of values:

	baseline	largest	smallest
a, Potential speed	20 m/h	30m/h	10m/h
b, % of route	80%	90%	70%

etc. for each other measure variable.

4, The MAP Processor forms "logical" (reality based) combinations of measure variable values. Examples of combinations of measure variable values for factor/task "Select proper route":

Combination 1: "potential speed" = 20m/h,
and "% of route" = 20%,
and (other variable values)

Combination 2: "potential speed" = 30m/h,
and "% of route" = 20%,
and (other variable values)

etc. for additional combinations.

5, Each "logical" combination of values of the measure variables describes an actual possible occurrence of the subject factor/task. For example, for the factor/task "Select proper route", the combination 1, above, describes the route selected by giving values to the measure

variables specified as descriptors of that factor/task, i.e., potential speed of movement = 20m/h, etc.

The authority provides a score representing the level of quality of the factor/task, e.g. the quality of the route selected, described by the combination of measure variable values. Example of authority's score of factor/task quality (example scale is 10.0 is best, 0.0 is worst):

Authority's score
for FACTOR/TASK

Combination 1: 8.0
Combination 2: 9.0
etc., for other combinations.

6, From the set of scored combinations, the MAP processor synthesizes a mathematical function of the measure variables that gives the same score for each combination as the authority provided. Example of a mathematical function for scoring the factor/task "select proper route":

"SELECT PROPER ROUTE": = $-40. + .075 \times (\text{potential speed m/h})$
+ $.183 \times (\% \text{ of route concealed})$ + other weighted variables.

Note: when mathematical functions for scoring the factor/tasks have been synthesized, a means for assigning numerical score values to factor/tasks i.e., quantifying factor/tasks, has been defined. Thus, numerical values assigned to a factor/task have physical meaning because a numerical value for a factor/task can be traced, via the mathematical function, to values of the measure variables which have physical meaning.

B, STEPS TO SYNTHESIZE QUALITY OF PERFORMANCE SCORE RULE (QPSR) FOR THE SPECIFIED MISSION FUNCTIONS

In the next set of steps (steps 7, 8, and 9), each combination of values of the factors/tasks describes unit performance on a mission. The authority scores each combination according to the quality of unit performance on the mission. Then, the processor synthesizes the quality of performance rule (QPSR) as the mathematical function of the factor/task scores that provides the same unit quality of performance score (QPS) for each combination as the authority provided.

7, Processor forms "logical" (reality based) combinations of factor/task scores. Examples of combinations of factor/task scores:

Combination 1
"Select proper route" score= 8.0, and
"Concealment during movement" score = 8.0,
and etc. (a score for each of the other factor/tasks.)

Combination 2
"select proper route" score= 9.0, and
"Concealment during movement" score = 8.0, and etc.
a score for each of the other factor/tasks.

etc. for other combinations of factor/task scores.

8, Authority scores each combination of "logical" factor/task values. Example of authority's Quality of Performance Score (QPS) for each combination of "logical" factor/task values (example scale is 10.0 is best, 0.0 is worst):

	SME's QPS
Combination 1:	8.0
Combination 2:	9.0
etc., for other combinations.	

9, Quality of Performance Score Rule (QPSR) is synthesized by the MAP processor from scored combinations of logical factor/task values, for example:

$$QPSR = -110. + 2.0 * (\text{Select proper route score}) + 1.0 * (\text{Proper cover/ concealment score}) + \text{etc.}, \text{ for the other weighted factor/tasks.}$$

STEPS TO ESTABLISH THRESHOLDS OF ACCEPTANCE FOR EACH VARIABLE

10. Authority identifies Minimum Acceptable Quality of Performance Score (MAQPS) for the specific mission function,

Example: $MAQPS > 8.5$

11, QPSR is then used to establish thresholds of acceptability for each factor/task from the minimum acceptable QPS for the specific mission function, Examples:

- a, Minimum acceptable value for "Select proper route" > 8 .
- b, Minimum acceptable value for "% of route concealed" > 7 .
- etc. for each other factor/task

12, Mathematical functions for scoring each factor/task are then used to establish thresholds of acceptability for each measure variable for the specific mission function. Examples:

- a, Minimum acceptable value for "Potential speed of movement" > 25 m/h.
- b, Minimum acceptable value for "Percent of route concealed" $> 80\%$
- etc. for each other variable.

13, Mathematical functions for scoring each critical man/machine/env. (CMME) parameter are then used to establish thresholds of acceptability for each CMME. Examples:

- a, Minimum acceptable value for "Slope" < 30 degrees
- b, Minimum acceptable value for "Max speed of vehicle" > 45 m/h
- etc. for each other CMME.

REFERENCE

Connelly, E. M., Performance measures for aircraft carrier landings as a function of aircraft dynamics. Paper presented at the Human Factors Society, 27th Annual Conference, Norfolk, Virginia, October 1983

MOMENT TO MOMENT PERFORMANCE ASSESSMENT OF MILITARY SYSTEMS

Edward Connelly

Communications Technology Applications, Inc.
7927 Jones Branch Drive, Suite 300
McLean, Virginia 22101
(703) 847-5710

Abstract

A unit Measure of Effectiveness (MOE) is a summary measure providing an effectiveness score for an overall mission. Directly, weapon system effectiveness can be measured by the weapon system's contribution to unit effectiveness. Indirectly, the system can also impact on unit effectiveness by limiting the performance of other systems. Thus, system performance can effect unit effectiveness both directly and indirectly.

By analogy, the concept of direct and indirect effects can be applied at a lower organizational level. System task performance can directly impact on system performance and also indirectly by limiting the performance of other system tasks. When assessing the impact of candidate systems designs, training programs, changes in doctrine, MOS staffing etc., the total impact of task performance must be known. A method for calculating the total impact i.e., both direct and indirect effects, is described here in terms of the system/task relationship; but the method applies equally to the unit/system relationship.

A Moment-to-Moment (MTM) performance measure provides an assessment of the performance of each system task during a mission. The assessment is the total (direct and indirect) impact of task performance on system performance, based on the system performance measure. The task performance assessment is calculated as the task is performed by taking into account the performance coupling of each task with that of other tasks. With a MTM measure, task performance is assessed according to its total impact on system performance and not by considering the task to be performed in isolation. The MTM measure can be used to accomplish the following: select tasks for training, establish task performance measures and levels of acceptable performance; evaluate proposed training programs, equipment designs, and procedures; and design system test and evaluation strategies. The paper describes the method of developing MTM measures as applied to team performance measures for military computerized systems and maneuvering units.

Method

Design of training programs requires analyses of the impact of task performance on overall unit/system effectiveness. Tasks are selected for training because: their performance is critical to mission success, the trainee population do not already possess the necessary skills, on-the-job training is not practical, and the use of job aids will not result in an adequate level of task performance. To make these determinations two types of measures must be available: (1) a quantitative measure for assessing overall unit/system effectiveness and (2) a means for assessing the impact of task performance on overall unit/system effectiveness.

When both types of measures are available, effectiveness based task selection is possible and design of the task training program can be effectiveness focused. Further, the measures can be used to design both post training tests as well as field tests.

Without a quantitative summary measure that can be relied on to assess overall unit/system effectiveness, the training program designer has great difficulty in:

1. Establishing a criterion level of effectiveness clearly identifying the acceptable level of overall unit/system effectiveness,
2. Communicating rules for assessing unit/system effectiveness or acceptable overall effectiveness criteria.

Without a means for measuring the impact of task performance on overall unit/system effectiveness, the training program designer has great difficulty in:

1. Identifying measures of task performance guaranteed to be relevant to unit/system effectiveness
2. Translating overall unit/system effectiveness criteria into consistent task performance criteria, and
3. Defending training strategy decisions in terms of their impact on unit/system effectiveness.

MOE Definition

A unit (or weapon system) Measure of Effectiveness (MOE) is a summary measure providing an overall unit effectiveness score for a mission. The MOE is composed of measures of the factors and tasks that are considered when assessing unit/system effectiveness. We refer to the factor and task measures as FT. Connolly (1987) discusses a theory of summary measure synthesis and gives an illustrative application.

Mission Task Coupling

Coupling between tasks occurs when the way one task is performed limits or enhances the possible performance of other tasks. Assessment of task performance must include both the direct and indirect effects on overall mission effectiveness. Direct measures of task performance are the portion of the MOE i.e., the FT, that are indicators of performance of that specific

task. This may, for instance, consist of the number of errors made while attempting to perform the task and time to perform the task, including time to correct any errors. If the MOE uses these ETAs, they are combined with ETAs from other tasks to produce the mission summary score. These are the direct effects of the task performance. Indirect effects consist of the impact of task performance on the performance of other tasks.

Some examples will illustrate the coupling effects referred to here. An MOE has been developed for a tank reconnaissance mission (Connelly 1987). According to the MOE, overall mission effectiveness is a function of the performance of multiple tasks. The first task is "Select Route". The quality of performance of that task impacts the performance of some other mission tasks which are as follows:

1. "Concealment during movement" (route selected may not permit concealment along total route).
2. "Visual Search" (Route selected may not permit visual search over all of assigned sector).
3. "Early acquisition of targets" (Route selected may limit visual search and cause a delay in target acquisition).
4. "Engagement and elimination of targets" (Route selected may delay acquisition of target and prevent engagement of targets).
5. "Report Contact" (Route selected may prevent target engagement resulting in a failure to report contact).

In practice, it is difficult to find a truly uncoupled task. The quality of task performance usually can impact performance of other tasks to some degree. For instance, when a radio operator is successful in a communication task by transferring the intended message without error, the task would seem to be uncoupled. If additional time is required to transfer the message and that additional time does not impact performance of other tasks, that time is simply added to the mission score according to the rules of the MOE.

However, a distorted message or the additional time used to transfer a message, may impact performance of other tasks, causing a coupling between the task performances.

These indirect effects are included in the summary MOE score as the direct effect of each other task, according to the rules of the MOE. However, the true assessment of a task performance must include both the direct and indirect effects. Decisions as to the criticality of a task to overall mission effectiveness, design of a training program, and post training and field testing should be based on the true total impact of task performance on mission effectiveness. A theory for total assessment of task performance along with an illustrative application are given in the subsequent paragraphs.

Moment-to-Moment (MTM) Performance Measures

A theory of MTM performance assessment (Connelly 1981) gives the rules for assessing the total impact (both direct and indirect effects) of task performance on unit effectiveness. The theory, which was developed from optimal control theory, recognizes that performance is limited both by machine factors and by human factors. Recognition that such limiting factors exist, whether or not they are explicitly known, leads to a measurement equation that permits evaluation of the effect of either instantaneous or of interval performance on the unit effectiveness for the mission.

The theory recognizes that a superior unit would achieve a specific mission with a high degree of effectiveness, as assessed by the MOE. The lack of a perfect effectiveness score, e.g. satisfying the mission orders with no casualties, no loss of equipment, use of a few resources, etc., reflects the existence of limiting factors which may not be explicitly known. Even though the limiting factors, such as limiting human factors, may not be explicitly known, they are known to exist and to be embedded in the performances of superior units. Thus, mission effectiveness accomplished by any unit can be assessed by comparing its effectiveness against with that of the superior units.

For instance, evaluation of a unit's overall effectiveness is accomplished by simply comparing its MOE score to that of the superior unit. But, our goal is to provide not only summary assessments but also total task performance assessments.

To describe the MTM theory, a concept must be introduced as follows. A model of a superior unit's performance is developed to define a reference performance. This reference is developed from a combination of data sources including field observation, simulation, performance test data, and judgment of subject matter experts. It is not necessary that this reference actually or accurately represent any actual unit or even the best possible performance. It is simply a reference we will use as a standard.

The theory employs a concept somewhat similar to the golf "par" system. If you are a par player, you expect a certain score when all 18 holes are completed. As you play each hole, you can assess your performance at each hole by comparing your actual number of strokes to the par, taking the difference. The value of par is the reference number of strokes specified for the hole.

In terms of the MTM theory, par for a golf course is an Expected Accumulative Score to the Objective (EASO), given that reference performance is achieved. If the EASO is computed before a task is performed and after the task is performed, a Difference in EASO (DEASO) is obtained. DEASO equals the value of FT for the task provided that reference performance is achieved. When assessing performance of a task performed by a unit being evaluated, the task FT score is compared to the DEASO. The comparison reveals the impact of task performance on the overall mission effectiveness. We cannot simply compare the task FT with that of the reference performance because of the task coupling effects. Thus, the rule for calculating MTM for task i is:

$$MTM_i = FT_i - DEASO_i$$

where $DEASO_i = EASO_b - EASO_a$
 $EASO_b$ is the EASO measured before task i is performed, and
 $EASO_a$ is the EASO measured after task i is performed.

Note that MTM_i is always zero for reference performance.

The analogy between the golf par system and the rule for scoring task performance is not exact. An important difference is that golf scores are not coupled from one hole to the next (except perhaps for the frustration carryover). However, task performances are often coupled. This means that the value of $EASO_b$ is a function of the way the task is performed.

An example may help to explain the process and demonstrate the benefits of MTM performance measures. Consider a mission illustrated in Figure 1, where the orders are to transit from point A to point D, as rapidly as possible. The terrain is such that route A,B,C,D is concealed from the enemy but points E and F are visible to the enemy. Obviously, the desired path uses points A,B,C, and D. However, we want to score any decision i.e., any path segment.

According to the MTM rule, reference task performance is established by determining reference FT for each possible path segment. These FT values are shown in Figure 2 by the numbers without circles around them. Thus, the move from point A to point E is scored as 2 FT units. The FT used in this mission is defined by the MOE and, for instance, might be travel time, or a combination of travel time and risk of detection by the enemy, etc. The Expected accumulative score to objective (EASO) is shown in Figure 3 by the circled numbers. To compute EASO, it is assumed that the reference performers would use the path that results in least overall EASO from the present point (state) to the objective. Thus, given the unit is at point E, the best next point on the path to the objective is F not B. Finally, the MTM measure is calculated for each possible path selection decision using the MTM rule given above. For simplicity of this description, it is assumed that the unit moves with the same speed as the reference unit so only the path selection will have a non-zero score. (Recall that all performances consistent with the reference receive a zero MTM score according to the MTM measure.)

As shown in figure 3, the MTM measure is a score for each possible path selection where the score value is the additional amount that will be added to the MOE due to the path selection. Since reference decisions always carry a zero MTM score, the score value assigned to other decisions reflects the severity of non-reference decisions. Severity is the additional amount added to the MOE because of the non-reference decision (or in general non-reference performance).

In summary, the MTM measure:

1. Scores each decision based on its total impact on overall mission effectiveness,
2. Clearly shows how to score each (decision) task as it occurs,

3. Reveals the states where correct (reference) performance is critical,
4. Shows states where training should be emphasized (because the penalties for errors at those states are large),
5. Shows where units should be tested (measure identifies states where reference performance is critical).

References:

Connelly, E.M., Comeau, R.F., & Steinheiser, F., (1981) Team performance measures for computerized systems (Final Tech. Report for ARI Contract No. MDA-903-79-C-0274).

Connelly, E. M., (1987) A Theory of Human Performance Assessment. Paper presented at the Human Factors Society 31st meeting New York.

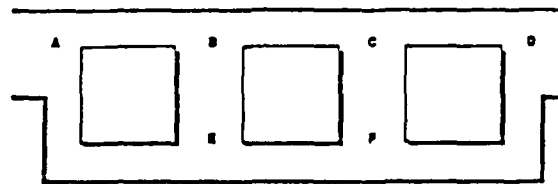


Figure 1. Unit States

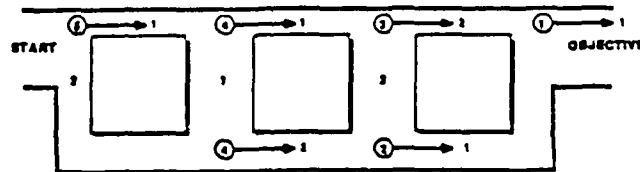


Figure 2. States Transition Scores and Expected Accumulative Scores to Objective (EASO)

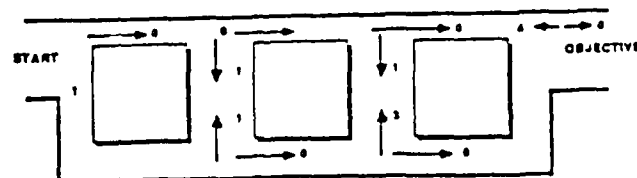


Figure 3. MTM Performance Measure

Diagnosing Trainee Learning Difficulties
Through Levels of Achievement Analysis

George M. Usova, Ph.D.
Office of Training Program Management
Internal Revenue Service
Washington, D.C.

Background. Recently, the Department of Labor commissioned the Hudson Institute to study and project trends for the National workforce in the year 2000. The final report, Workforce 2000, depicts demographic trends that may have negative consequences for millions of individuals and for our society as a whole unless National policies toward employment, education, and training are modified.

According to Packard (1988), co-author of Workforce 2000, the education of America's workforce is inadequate to meet the demands of the next century. He states,

Even as the workplace becomes more complex, our hard-pressed inner-city schools are responsible for educating a growing fraction of tomorrow's labor force. Each year 700,000 young people drop out of high school and an equal number graduate without functional literacy. Add to that a million new working-age immigrants, and we have almost 2.5 million persons entering our complex economy annually with limited language and work skills.

In 1986, minorities accounted for about 21 percent of the jobs in the American workforce of 115 million. Between 1986 and the year 2000, the number of jobs will increase by 21 million - and an astonishing 57 percent of those additional jobs will be filled by minorities. Yet if present trends continue, disproportionate numbers of those workers will lack the skills needed to do the job properly. In other words, unskilled minorities are a growing fraction of the workforce and unless their abilities are upgraded, the nation's overall skill level will not be sufficient for tomorrow's economy.

Current Status. The Educational Testing Service recently conducted a National Assessment of Educational Progress (NAEP) survey of 3,600 21-to-25-year-olds representing the entire population of that age bracket who lived in households. It came up with these discouraging figures.

- o Only 80 percent of whites, 60 percent of Hispanics and 40 percent of blacks could locate two items of information in a sports article.
- o Only 65 percent of whites, 35 percent of Hispanics and 20 percent blacks could follow directions from one location to another using a map.

- o Only 80 percent of whites, 60 percent of Hispanics and 40 percent of blacks could enter and calculate a checkbook balance.

Skill Forecast. The "Workforce 2000" report anticipates a gain of 25 million workers from 1984-2000. A Department of Labor standard rates the reading and communication skills that will be required for projected job areas on a scale of one to six. (A level 1 job requires a reading vocabulary of 2,500 words and the ability to write a simple sentence. Level 6 job-holders must use technical journals, financial reports and legal documents.)

Of the net increase of about 25 million jobs expected to be created between 1984 and 2000, "Workforce 2000" projects that:

- o Nearly 40 percent, or 10 million jobs, will be professional or technical positions requiring language skills of level 4 or better.
- o Another 58 percent, or 15 million jobs, will be marketing and sales, administrative, services, supervisor and similar positions requiring skills levels between 2.5 and 3.9.
- o Only 2 percent, or about a half a million jobs, will require language skills less than 2.5.

"Workforce 2000" forecasts that the average skill level required for these new jobs will be 3.6. But NAEP's estimate of the skills of young white, black, and Hispanic adults averages only 2.6. Thus 25 million new workers will have to improve from 2.6 to 3.6 on the Labor Department scale - a 40 percent increase.

Levels of Achievement. Given the onset of a declining workforce in terms of numbers and skill levels, it is incumbent upon the training organization to diagnose and assist trainees in overcoming learning difficulties. Trainers should be equipped with instructional diagnostic techniques that assess learner abilities or difficulties prior to the implementation of training. One such technique is the Levels of Achievement analysis that assesses trainee ability to read and understand their present training materials. The diagnosis yields information for making appropriate decisions for instructional enrichment, redirection, or remediation.

The Levels of Achievement analysis yields data for trainees in three category levels: (1) Independent (2) Instructional and (3) Frustration. Each level contains criteria from which trainee standards of reading and comprehension performance are measured.

The technique first requires the trainer to identify a representative 300-400 word excerpt from the materials. From this passage, the trainer then constructs 10 comprehension questions. Now, the trainee is asked to read the passage aloud. During this process, the trainer (examiner) records any oral reading errors (substitutions, mispronunciations,

omissions, repetitions, etc.) and calculates an oral reading error percentage, e.g., 10 oral reading errors in a 200 word passage is a five percent error rate. The trainee is finally asked to answer the 10 comprehension questions from memory; similarly, the examiner calculates a total comprehension score and records it as a percentage. The trainee's performance is then matched against the following criteria established by Betts (1970) to determine the trainee's Level of Achievement category.

	Oral Reading Accuracy	Comprehension Accuracy
Independent	99%	90%
Instructional	95%	75%
Frustration	<90%	<50%

Given the trainee's performance and consequent Level of Achievement category, the trainers can judge the level of success the trainee will have with the materials themselves, and make appropriate decisions on enrichment or remediation.

MEASURES OF EFFECTIVENESS: Increasing Their Sensitivity to Personnel Characteristics

Mark Y. Czarnolewski
John L. Miles, Jr.

*U.S. Army Research Institute
Alexandria, Virginia*

It is axiomatic that one does not commit resources to an engineering program with no discernible means of evaluating return on investment (Miles & Quinkert, 1988). Normally what that means is some form of measurement within relevant parameters. Those who evaluate weapon system performance generally select the parameters of effectiveness, availability, and reliability (Grubbs, 1979). Within these parameters are both physical phenomena and performance characteristics which can be measured directly. Instrumentation and methods for making such measurements have been part of the Army and industry testing tradition for years.

But direct measurements of these parameters, while undeniably helpful, do not provide all of the information desired by decision-makers. There seems to be an increasing realization in the Army R&D community that other factors—some of them far less susceptible to direct measurement—can play a significant role in the “value” that a new weapon system can bring to the battlefield. Many of these factors are related to the personnel who operate and maintain the weapon system and to the leaders of those personnel. Lowry and Seaver (1986) have proposed a methodology for accounting for the measured soldier performance of critical operations and maintenance tasks, and Borman, et al. (1987) have described a model of fifteen psychological parameters of individual soldiers. The question faced by those who are asked to test and evaluate a complex new weapon system today is, “How can we put all these factors together into some meaningful metric?”

Two primary factors reflecting human performance in missile systems in general, and in TOW gunnery in particular, are detection and identification of a target and control of the TOW missile once it is fired (Cartner, et al., 1985). One way of operationalizing these characteristics is by use of a measure of effectiveness (MOE). MOEs avoid the potentially fuzzy logic of assigning military value to direct measurements (e.g., 20 pounds is “good”, but 28 is “marginal”) and instead include in their design some scheme for incorporating military value which is both rational and quantitative.

A number of schemes for building MOEs can be found in the literature for combining both human and equipment performance into a metric of military value. Among the best is a report by Tiedemann and Young (1970), which explains the logic of making mathematical combinations of unlike measures. Following a similar sort of approach, O’Keefe and Guerrier (1988) present an MOE for analyzing the performance of TOW-2 gunners during a recent field test at Fort Lewis.

In the Fort Lewis test, the planners realized that, while hit probability (H) is always likely to be the first performance indicator examined, there may be a sharper and more useful measure of gunner performance. They combined the H data with their second performance measure (time to acquire the target) in such a way that a “fast hit” received a numerically greater score than an “eventual hit.” The formula for their MOE is:

$$MOE = H \times \frac{K-T}{K-1} \quad (1)$$

where: H = hit probability; T = target acquisition time, given a hit; and K = a constant larger than any T .

While that MOE seems to offer system evaluators a better metric of effectiveness than H alone, it does not make full use of the time parameter. The MOE in Formula 1 only considers time when there is a hit; it does not capture the component of time that reflects the processing related to a miss. Consequently, it precludes full determination of the interplay between a soldier's detection time and his control of the missile's flight toward the target.

We propose that the following characteristics and methods be incorporated into an MOE for TOW gunnery performance: (1) Direct measurement of the sequential aspects of the task; first, detection time and, second, behaviors that keep the missile on target (c.f., Logan, 1985); (2) Detection time of hits and misses to capture the processing of target detection and identification, as well as the interplay between detection time and eventual accuracy performance (Pachella, 1974); and (3) Identifying consistency of individual differences by determining those stimulus conditions affecting performance and determining the relationships among the conditions (thereby allowing for greater homogeneity and—therefore—greater reliability of performance measurement) (Nunnally, 1978).

Our intent is to incorporate into the MOE certain psychometric principles, experimental psychology models that consider speed and accuracy, as well as the sequential characteristics of the TOW gunnery task. Our primary hypothesis is that by incorporating these characteristics, one is more likely to identify consistency in individual differences of performance across stimulus conditions than when one does not consider these characteristics.

Method

Eighty-five soldiers from the 9th Infantry Division (Motorized) at Fort Lewis, Washington, participated in this research project. Procedural problems resulted in only 51 soldiers completing a sufficient number of trials for the criterion data that were analyzed for this paper. The soldiers, who had above average GT scores ($\bar{x} = 104.5$, $\sigma = 12.2$), also had varying degrees of experience with firing a real TOW2 missile and with practice on the DX164 simulator of the TOW2 (O'Keefe & Guerrier, 1988).

The DX164 is a training device for the TOW2 missile system which trains gunnery and tracking skills and is designed to fire simulated TOW2 missiles against real and simulated targets. It provides feedback on acquisition time (i.e., the time between a target's appearance and the soldier's firing a simulated missile), hits and misses of the target, and other dependent measures. It is used at the Anti-Armor Theater (AAT) at Fort Lewis for TOW2 training of 9th Infantry Division personnel (SFC Guillen, Personal Communication, October, 1987).

The primary research effort described in O'Keefe and Guerrier was a test and evaluation of the overall training effectiveness of the DX164 and three specific training techniques. The design of the primary effort may be described as a 3 (training technique) X 3 (stimulus condition) X 2 (testing session) repeated measures, with training technique being a grouping factor and stimulus condition and practice representing repeated measures.

Each firing point represented a qualitatively different stimulus condition. At Firing Point 1, all the soldiers wore all the gear required for nuclear, biological and chemical (NBC) conditions. Firing Point 2 did not require NBC gear and had the shortest and least occluded gun-target ranges. Firing Point 3 also did not require NBC gear, but contained targets at the greatest ranges and with the most visual occlusion. This paper analyzes only a portion of the data presented in O'Keefe and Guerrier; specifically, four engagements at each firing point where the gunners engaged real targets during daylight conditions. Our data are further limited by including only post-training performance. Gunners observed the target through the daysight (even though doctrine prefers use of the nightsight to overcome effects of some obscurants). Order of firing point was balanced across gunners. Further details of the stimulus conditions at each firing point may be found in the O'Keefe and Guerrier report.

Results

Table 1 contains descriptive statistics of acquisition times and hit probability for each firing point. As expected, Firing Point 2, which had the shortest ranges and least demanding conditions, also had the quickest average acquisition times and highest average hit probability.

Table 1
Means and Standard Deviations for Acquisition Time (Seconds) and Hit Probabilities of Each Firing Point

Firing Point	Acquisition Time		Hit Probability	
	<u>\bar{x}</u>	<u>S.D.</u>	<u>\bar{x}</u>	<u>S.D.</u>
1	10.0	3.8	86.0	21.6
2	5.3	0.8	95.9	12.0
3	9.9	1.6	82.0	17.8

Regression equations allowed us to construct parameters that reflect the interplay between detection time and hit probability. Of primary importance were parameters that capture the sequential aspects of the task, as well as those aspects of performance that are based on the behaviors accounting for differences in hit probability. For example, Component B in Figure 1 is operationalized by the residual ($H - \hat{H}$) based on the regression equations.

The result is :

$$MOE = (H - \hat{H}) \times \hat{T} \quad (2)$$

where (for each firing point):

$H - \hat{H}$ = that portion of hit probability that is exclusive of performance predicted by acquisition time, and

\hat{T} = that portion of the acquisition time measure that is predicted (or constrained) by one's hit probability.

Compared to Formula (1), which employs only acquisition times associated with a hit for computing an MOE, the MOE in Formula (2) includes those times where there may have been a miss. Cohen and Cohen's (1983) chapter on causal models provides examples that show how residuals and predicted values can be operationalized in the analysis of correlated data.

One employs the sign of the correlation between $(H - \hat{H})$ and \hat{T} to combine each parameter into an MOE for each firing point. A positive correlation indicates that higher acquisition times (\hat{T}) are likely to be related to more accurate missile controlling behaviors, $(H - \hat{H})$; consequently, higher \hat{T} times are given higher standardized scores when the two parameters correlate positively with each other. Conversely, for a negative correlation between these parameters, higher \hat{T} times are given lower standardized scores when the two parameters correlate negatively with each other. Accuracy scores are standardized as well, with the higher the score, the better the performance. One employs Formula 2 in this context when computing an MOE for a firing point.

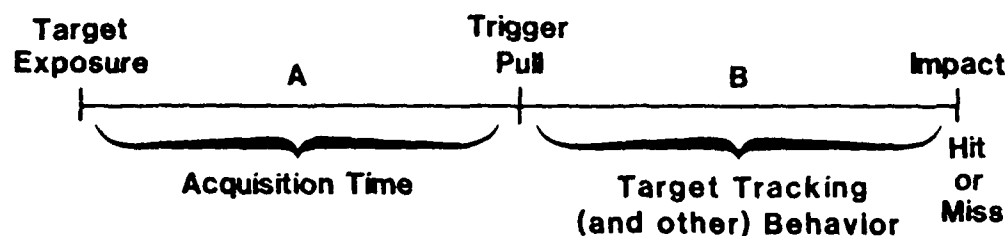


Figure 1. Diagram providing overview of the TOW gunner's task.

Finally, to combine each firing point's MOE to form a composite MOE, one can consider the sign of the intercorrelations among the MOEs of each firing point. Firing Point 3's MOE correlated negatively with Firing Point 1 and 2's MOE ($r = -.28, p < .05$ and $r = -.24, p > .05$, respectively); but Firing Points 1 and 2 correlated positively with each other ($r = .24, p > .05$).

To evaluate the efficacy of the proposed algorithm represented by Formula 2 to the O'Keefe and Guerrier algorithm using Formula 1, an "intermediately sensitive" MOE was evaluated as well. That is, a separate MOE using Formula (1) was calculated for each firing point. The measures taken at Firing Point 3 are given a negative weight when computing this MOE, because those measures correlate negatively with the MOEs for the other two firing points. It is hypothesized that the order of MOE sensitivity to measuring consistency in individual differences will be $((H - \hat{H}) \times \hat{T})$; Hit Probability x Acquisition Time (given a hit) per firing point; Hit Probability x Acquisition Time (given a hit). The Composite MOEs that reflect these three different models are labelled MOE3, MOE2, MOE1, respectively.

Table 2 contains the correlations of the three MOE composites with some of the tracking and spatial tests previously found to predict TOW gunnery performance with an M-70 simulator at Advanced Individual Training (AIT) (Grafton, *et al.*, 1988). Correlations of these tests with that portion of the data in the present study with which these tests most correlate are shown in Table 2 as well. The spatial tests are said to measure the abilities of mental rotation (Orientation Test), scanning (Maze Test), and detection of features to reassemble objects (Assembling Objects Test). The tracking tests look at the ability to rendezvous with a target moving along a pre-defined path, and to remain on top of the target as it moves along the path.

Table 2

Correlations of Alternative MOEs with Some Spatial and Tracking Tests ¹

MOE Composites:	Spatial		Assembling Objects	Tracking	
	Orientation	Maze		One- Handed	Two- Handed
MOE 1	.28	.32*	.05	-.19	-.20
MOE 2	.42**	.33*	.34*	-.20	-.15
MOE 3	.52***	.50***	.36*	-.27	-.23
Firing Point 1					
H	.53***	.33*	.14	-.23	-.19
MOE 2	.56***	.35*	.15	-.35*	-.33*
H - \hat{H}	.64***	.46***	.31*	-.39**	-.33*
MOE 3	.66***	.46***	.28	-.32*	-.24

Note: See text for definitions of MOE 1, MOE 2, and MOE 3 for both MOE Composites and for an MOE for a specific firing point.

* $p < .05$

** $p < .01$

*** $p < .001$

Table 2 shows that MOE3 was most sensitive to detecting consistency in individual differences across qualitatively different stimulus conditions, as represented by the three firing points. The model represented by MOE2 was somewhat less sensitive, and MOE1 was least sensitive to detecting consistency in individual differences across qualitatively different stimulus conditions. The MOE3 model was more sensitive than the MOE2 model, despite the fact that Firing Point 3 was not given a negative weight in MOE3, but was given a negative weight in MOE2. For Firing Point 1, accuracy exclusive of detection time, ($H - \hat{H}$), was more sensitive to individual differences than accuracy (H), and MOE3 appeared more sensitive to spatial test performance, but MOE2 more sensitive to tracking test performance. The primary hypothesis was, thus, confirmed for testing MOE sensitivity to individual differences across stimulus conditions, but not for one of the more difficult stimulus conditions.

Conclusion

Measures of effectiveness are widely used because they systematically add "military value" to what are essentially engineering data. This paper has presented one way in which an MOE can also

¹ NOTE: For the spatial tests, the larger the number of correct responses, the higher the score; for the tracking tests, the larger the off-track deviation score, the poorer the performance. Therefore, correlations with DX164 performance measures should be expected to be positive with the spatial test scores and negative with the tracking test scores.

be tuned to be sensitive to gunner characteristics. As the Army's MANPRINT program continues to demonstrate that soldier aptitude plays an increasingly important role in the battlefield performance of high-technology systems, test and evaluation personnel should insure that the metrics by which they infer value have also been examined for their sensitivity to the personnel characteristics associated with high-value performance.

References

- Borman, W.C., Motowidlo, S.J., Rose, S.R., and Hanser, L.M. (1987). *Development of A Model of Soldier Effectiveness* (Technical Report 741). Alexandria, VA: U. S. Army Research Institute.
- Cartner, J.A., Strasel, H.C., Evans, K.L., Heller, F.H., and Tierney, T.J. (1985). *TOW Gunner Selection* (ARI Research Note 85-72). Alexandria, VA: U.S. Army Research Institute.
- Cohen, J. and Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. (2d Ed.) Hillsdale, NJ: Lawrence Erlbaum Assoc.
- Grafton, F.C., Czarnolewski, M.Y., and Smith, E.P. (1988). *Relationships between Project A psychomotor and spatial tests and TOW2 gunnery performance: A preliminary investigation* (Working Paper). Alexandria, VA: U.S. Army Research Institute. In process.
- Grubbs, Frank E. (1979). *Army Weapon Systems Analysis, Part One* (DARCOM Pamphlet 706-101). Alexandria, VA: U.S. Army Materiel Command.
- Logan, G.D. (1985). Skill and automoticity: Relations, implications, and future directions. *Canadian Journal of Psychology*, 39, 367-386.
- Lowry, John C. and Seaver, David A. (In Press). *Handbook for Quantitative Analysis of MANPRINT Considerations in Army Systems* (Research Product 88-15). Alexandria, VA: U.S. Army Research Institute.
- Miles, John L. Jr., and Quinkert, Kathleen A. (1988). *Can MANPRINT and "Fightability" Be Tested? (Reprise)* Unpublished Manuscript.
- Nunnally, J. (1978). *Psychometric theory*. (2d Ed.) New York: McGraw Hill.
- O'Keefe, Timothy J. and Guerrier, Jose H. (1988). *Training Effectiveness Evaluation of the DX164 TOW2 Missile Firing Simulator* (ARI Technical Report). Alexandria, VA: Allen Corporation of America Technical Report OPM-85-75/W.O. 375-120.
- Pachella, R.G. (1974). The interpretation of reaction time in information processing research. In B.H. Kantowitz (Ed.) *Human information processing: Tutorials in performance and cognition*. Potomac, MD: Erlbaum.
- Tiedemann, A.F. Jr. and Young, R. Bruce (1970). *Index of Proximity: A Technique for Scoring Suppressive Fire* (Engineering Report 6419). Baltimore, MD: AAI Corporation. [AD 723416]

Toward a Diagnostic Operator Performance Assessment Scheme

Dr. Laurel Allender

US Army Research Institute Fort Bliss Field Unit

Mr. Bryan E. Brett

Science Applications International Corporation

Today's automated weapon systems and tomorrow's automated command and control systems provide a vehicle for automated performance assessment. The system computer, in addition to its primary function, can collect, analyze and report operator performance data for the researcher, the trainer or the operator. However, the sheer power of automation alone does not mean that a performance assessment capability is adequate. Vreuls and Obermayer (1985) report on the status and fundamental problems of automated performance assessment capabilities. They state that many of the existing capabilities are woefully inadequate and largely useless.

Three particular problems identified by Vreuls and Obermayer are addressed through the performance assessment capability presented in this paper. (1) Performance assessment of automated systems must be descriptive of both system performance and operator performance. The assessment of system performance must consider the combined contributions of hardware, software and operator performance whereas the assessment of operator performance must consider only the unique contribution of the operator. (2) A performance assessment capability must capture performance of decision-driven tasks. An operator's job on a sophisticated weapon or command and control system is not strictly one of rote, procedural steps. He or she employs strategies, exercises judgment and makes decisions. Thus, a specific action is not necessarily right or wrong, but must be considered with respect to the context and the mission outcome. (3) A performance assessment capability must be diagnostic. The end-of-mission scores typically collected, while useful, do not provide answers to the questions "what did the operator do wrong?" or "what did the operator do right?" Specifics must be identified in order to effect system, operator or training changes.

The Performance Assessment Capability (PAC)

The Army Research Institute (ARI) Fort Bliss Field Unit, under the High Altitude Air Defense-Console Operator Performance (HAAD-COP) research program, has developed a performance assessment capability based on an earlier conceptual foundation laid by Hawley and his colleagues (Hawley, Brett, & Chapman, 1982; Hawley, Howard, & Martellaro, 1982) and on a partial implementation as reported by Allender (1987). The current PAC addresses all three problems: (1) The PAC is descriptive of both system and operator performance. It comprises four levels, each successive level telescoping out to more specific detail. (2) The PAC captures decision driven tasks. It is not derived from a traditional task analytic approach. It is mission based so that individual actions are not considered right or wrong, but are considered in context as they contribute to mission outcome. (3) The PAC supports performance diagnosis. Specific actions are directly linked to mission outcome and can be used to diagnose performance problems and successes.

The PAC was implemented on the Patriot Tactical Operations Simulator (PTOS) operated by the Directorate of Combat Developments at the US Army Air Defense Artillery School at Fort Bliss, Texas. The PTOS is a real-time, high-fidelity mission simulator of the firing battery and battalion operator consoles of the Patriot missile system, the Army's fully-automated high altitude air defense weapon system. At both the battery and battalion levels, the operator responsibilities are divided into two functions. One operator is designated the friendly protector, the other the weapons controller. The friendly protector is responsible for verify-

ing and overriding system assigned aircraft identifications and for preventing engagement of friendly aircraft. The weapons controller initiates aircraft engagements.

At a Patriot console (see Figure 1.), the operator is provided with a display that graphically presents elements of the air defense situation (e.g., location of assets, corridors, identification volumes) and symbols indicating the current identification of each aircraft (friend, unknown, hostile). In addition to the graphical display, tables of numeric and text information can be shown. Located on the console around the display are numerous switches the operator uses to obtain information to support decision making or to implement a decision (e.g., place an identification on a aircraft, override an engagement).

The structure of the PAC and the specific measures used on the PTOS are shown in Figure 2. The highest level is the mission level. The measures are based on combined hardware, software and operator contributions. Defense of assets is the percent of assets successfully defended by the total system. Attrition is the percent of hostile aircraft eligible for engagement that were successfully killed. Friendly protection is the opposite of fratricide. It is computed as 100% minus the percent of aircraft scripted as friends that were killed. Resource conservation is the percent of missiles fired that were not wasted. Values less than 100% for any of the mission-level measures point to a performance deficit.

The level below mission-level measures is the function level. Function-level measures summarize operator performance according to the weapons controller or the friendly protector task allocation and as such measure unique operator contribution to overall system performance. For example, looking at the friendly protector, some of the measures are the percent of aircraft identified by the operator, the percent of those identified correctly according to the scripting of the scenario, the average delay to identification, and the percent identified late. This last measure deserves a special comment as it is tied to a central concept in the PAC--the task performance window. For each task there is a window of time during which performance of the task will be optimized. If a task is performed before the beginning of its window, not all information is available and the task may have to be repeated. If a task is performed after the end of the window, some undesirable event will occur--a friendly engagement, a wasted missile, an asset penetration.

The third level of diagnostics is the task level. Patriot operator functions are not characterized by rigid steps. There are, however, certain actions or tasks that must occur to accomplish function-level performance. Examples of such actions are percent of aircraft hooked (or selected for further action) and average delay to acknowledge alerts. The lowest, or most detailed, level of diagnostics is the performance timeline. It is a detailed description of operator actions for each aircraft in the scenario in the context of other critical aircraft and system events.

Experiment

Method and Procedures. The PAC was exercised recently in a baseline investigation of Patriot friendly protector performance conducted with the PTOS. The test operators were 24 trained Patriot operators from Fort Bliss, Texas. Each operator was familiarized with the simulation equipment and the tactical situation

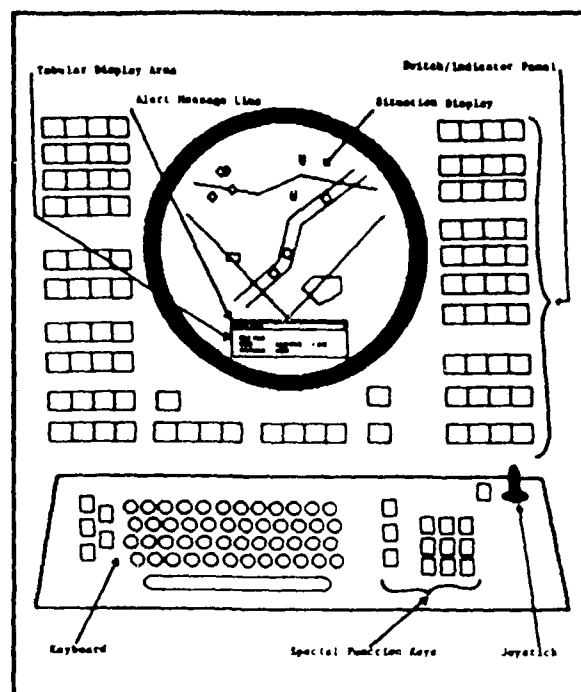


Figure 1. Illustration of Patriot operator console (TM-9-1430-600-10-1, 1983)

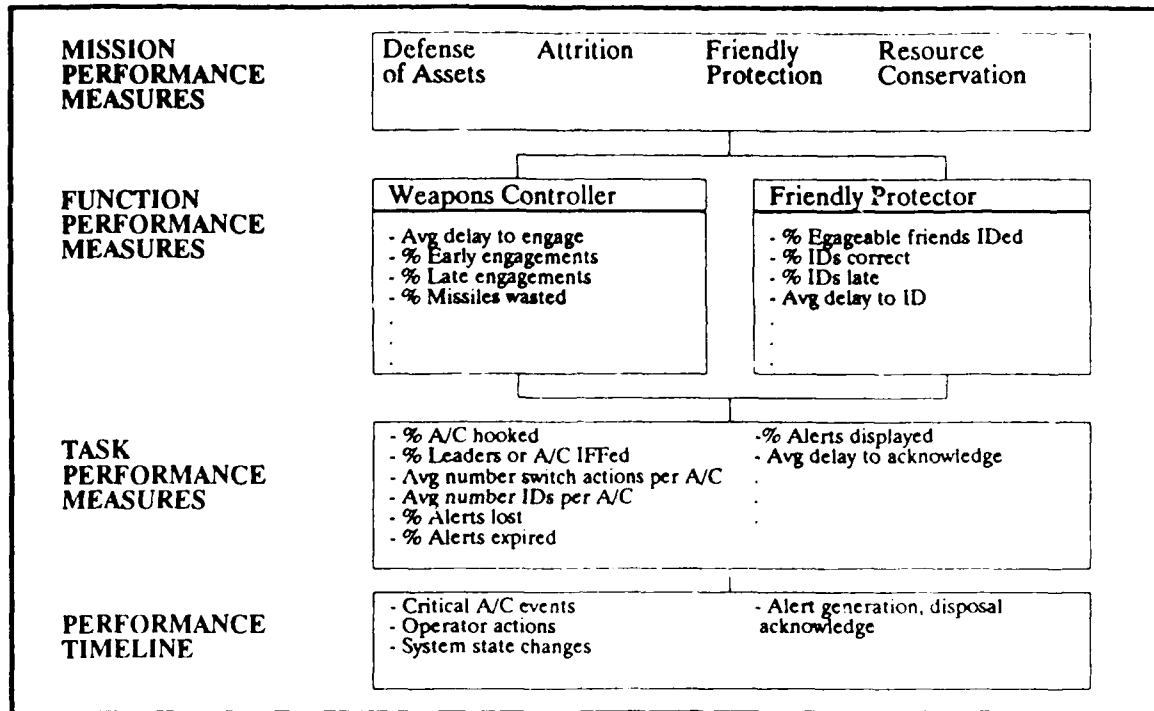


Figure 2. Structure of HAAD-COP PAC.

during a one half hour training scenario. The twelve test scenarios were presented to each subject in a different random order. The testing session lasted three hours.

Adhering to standard operating procedure, the system software automatically assigned an identification to each aircraft based on various events. Recall that the friendly protector's function is to verify and override the system-assigned identifications and to prevent engagements of friendly aircraft. To ensure that identification override was required throughout the scenarios, the operator was provided with information on some aircraft that was not available to the system software. Specifically, the operators were given formation proximity criteria so that they could impose flight leader identifications on all aircraft meeting the criteria. Without the operator-assigned identification, the system would identify only the flight leader as a friend, not the wingmen. The function of weapons controller was performed automatically by the PTOS software.

Results. Analysis of the data from the experiment is still ongoing. The results presented here are based on data from ten test operators. The mission-level measures are presented in Table 1. In the case of each of the four measures, performance is less than 100%. The two measures most directly affected by the friendly protector function are friendly protection and resource conservation. The friendly protection score indicates that 14% of the friendly aircraft were killed. The resource conservation score indicates that over 20% of the missiles were wasted. In order to diagnose the causes of the performance deficiencies, the other levels of performance assessment must be examined.

Defense of Assets	Attrition	Friendly Protection	Resource Conservation
94.29	91.97	85.46	78.43

Table 1. Mission-level measures averaged across all 12 scenarios. N = 10 for all measures. All measures are reported as percents.

The function-level measures showing the friendly protector's unique contribution to the mission measures are presented in Table 2. Of the total aircraft that were scripted to be friends in each scenario, only 46.36% were identified by the operator.

All Friendly Aircraft		Engageable Friendly Aircraft			
% IDd	% Engageable	% IDd	% ID Correct	Avg Delay to ID	% IDs Late
46.36	57.98	54.88	97.60	84.42	47.46

Table 2. Function-level measures averaged across all 12 scenarios. N = 10 for all measures. Average delay to identify is in seconds.

Of those, an average of 57.98% were at risk, that is, were engageable by the system unless the operator intervened and overrode the system-assigned identification; however, only 54.88% of those were identified by the operator, although a high number, 97.60%, were identified correctly. The identifications of engageable friends were made, on average, 84.42 seconds after the beginning of the task performance window, almost a minute and a half after all the information was available upon which to make a correct identification. In turn, that minute and a half resulted in 47.46% of the identifications occurring after the end of the task window. In other words, in all likelihood, 47.46% of the identifications took place after an engagement had been initiated. To summarize so far, the mission-level performance deficiencies are linked to two function-level deficiencies: (1) just more than one-half of the at-risk friendly aircraft were identified by the operator, but (2) even of those that were, nearly half of the aircraft were identified after the critical end of the performance window.

The task-level measures shown in Table 3, further diagnose the performance deficiencies seen thus far. Fully 91.75% of the friendly aircraft were hooked, that is, selected in order to examine the aircraft history information, or to implement a decision-either to impose an identification or to override an engagement. However, the function-level measures show that only 46.36 of all friends were identified by the operator. Thus, some aircraft were hooked, but were not identified by the operator, resulting in friendly aircraft being killed by the system. Other aircraft were hooked and identified only in the process of stopping an engagement, resulting in a saved friend but a wasted missile. Also, in some cases, the operator imposed an identification more than once on a given aircraft.

% Friends Hooked	% Friendly Engagement Overrides	Avg Number IDs per A/C (Friends & Hostiles)
91.75	50.49	1.18

Table 3. Task-level measures averaged across 12 scenarios. N = 10 for all measures.

Looking at the most detailed level of analysis in the PAC, the performance timeline shown in Figure 3, a typical sequence of events unfolds. For the two aircraft shown, a flight leader and a wingman, the window begins at 300 seconds into the scenario and ends at 455 seconds. For the top aircraft, the flight leader, the operator imposes the identification of "friend" at 450 seconds; however, it is only when the timeline for the bottom aircraft, the wingman, is examined that the stimulus for the identification of the flight leader becomes apparent. The flight leader is carrying an identification of "unknown" based on a positive Mode 3 Identify Friend or Foe return combined with negative airspace volume violations; the wingman is carrying an identification of hostile based on only the negative volume violations. At 445 seconds the wingman is engaged by the system, prompting the operator to examine the formation more closely. An engagement, then, is the stimulus for the operator identification of both friendly aircraft. Both aircraft were protected, but a missile was wasted. The operator, in this case, simply failed to examine the formation in time.

Another source of performance deficiency is evidenced in detailed performance timelines where aircraft in a formation are hooked in succession, but one aircraft is identified repeatedly, while the other two or three are not identified at all. Here the problem can be traced to difficulty with the hooking procedure itself and to

the display of information. Operators were apparently confused about which aircraft was hooked due to unclear feedback either on the graphical display or on the tabular display.

Discussion

The diagnosis of the system performance deficiencies seen here with the function of the friendly protector performed by trained operators on the PTOS points to several remedies. Problems with the mechanics of the hooking procedure and the display of feedback could be addressed through fairly straightforward software changes or through a training emphasis. Problems with late identifications could also be addressed to a degree by training operators to make decisions earlier, perhaps by training the notion of a task performance window directly. At the same time there is an risk of increased workload if decisions are made too early and have to be repeated later. Finally, problems with the complete failure to identify some aircraft at all may also be addressed to some degree through training, but an arena for serious investigation is the enhancement of system software to handle decisions such as formation identification and to prompt operators to those cases requiring further resolution.

In conclusion, the PAC described here addresses the three major problems of many existing automated performance assessment capabilities. It is descriptive of both system and operator performance; it captures the flow of decision-driven tasks; and it is diagnostic. The results discussed here show the applicability of the PAC as research tool and as a means for recommending training and system solutions to problems. Currently, the PAC is being studied for applicability as a trainer's aid and in supplying operator training feedback directly on a training device.

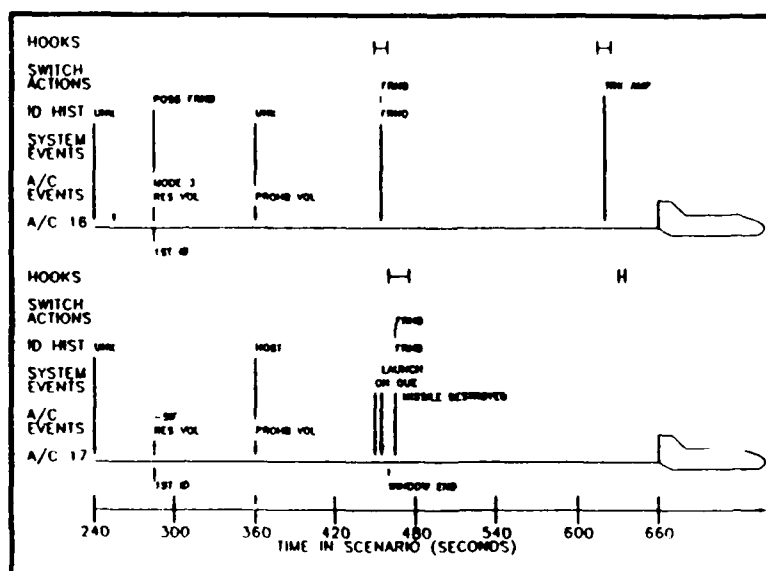


Figure 3. Performance timeline illustrating a late identification..

REFERENCES

- Allender, L. (1987). An evaluation of the usefulness of color in the Patriot display. Arlington, VA: US Army Research Institute Working Paper FB 87-01.
- Hawley, J. K., Brett, B. E., & Chapman, W. A. (1982). Optimizing operator performance on advanced training simulators: Specifications for operator performance assessment and workload quantification. Valencia, PA: Applied Science Associates.
- Hawley, J. K., Howard, C. W., & Martellaro, A. J. (1982). Optimizing operator performance on advanced training simulators: Preliminary development of a performance assessment and modeling capability. Arlington, VA: US Army Research Institute Technical Report 573.
- TM 9-1430-600-10-1 Operator's manual: Engagement Control Station, Guided Missile, Truck Mounted: AN/MSQ-104 (1983). Washington, D.C.: Department of the Army.
- Vreuls, D., & Obermayer, R. W. (1985). Human-system performance measurement in training simulators. Human Factors, 27, 241-250.

A Comprehensive Method for Evaluation of Critical Military Training Programs

Jeffrey A. Cantor
Professor, Lehman College CUNY
& Consultant, DDL OMNI Engineering
Chantilly, VA

C. Lee Walker
Manager Evaluation
DDL OMNI Engineering

This paper describes a comprehensive approach to military training evaluation based upon systematic analyses of program objectives. It permits identification of discrepancies between training program objectives and components and program outcomes. This has proven useful in evaluation of critical technical operations training where multiple data sources must be analyzed against specific baseline standards to determine program congruence and personnel competence.

The Discrepancy-Based Total Training Program Evaluation was developed in two environments, the US Navy's strategic programs and the civilian nuclear power industry. Within the Navy the critical nature of technical training is well recognized. In the civilian sector, after the Three Mile Island accident, a more rigorous analysis of employee performance and plant training became necessary to ensure worker competence and organizational compliance with governmental guidelines. A methodology was needed which would provide formative prescriptive analyses of the training program and its outcomes, as well as summative findings of the program's effects in meeting its goals. The process also had to be capable of use by external teams for periodic accreditation (summative) reviews.

A Conceptual Framework:

The objective of military training program evaluation is to provide a systematic process for collecting, reviewing, and analyzing critical personnel performance data in accordance with a service and/or system/command's engineering system and organizational/procedural baseline specifications. To meet this objective the tool must provide a process to analyze the command and/or system's training standard baseline against the service's doctrine and policy, and ultimately against critical personnel performance outcomes. The process must provide formative findings for immediate attention and summative findings for periodic review and reporting to external service/ governmental and policy making organizations. It must be usable as an ongoing command and/or system's evaluation process.

The primary purpose of training program evaluation is to determine whether to maintain, improve, or terminate a program. The Discrepancy-Based Total Training Program Evaluation process incorporates discrepancy analyses processes to facilitate: (a) identification of program baseline standards; (b) determination of whether a discrepancy exists between the training program constructs and the baseline standards governing that program; and (c) use of "discrepancy" information to identify the weaknesses of the program. Discrepancy analysis accomplishes formative and

summative levels of evaluation permitting immediate outcome reporting as well as long range findings. The overall evaluation specifies: a) discrepancies between best commercial practices for instructional systems design and development and the service and system and/or command's baseline training standard; b) discrepancies between engineering and operational requirements and the command's training program; and, c) discrepancies between the command and/or system's baseline training standard and the program's training outcomes.

Discrepancy analysis allows stipulation of program outcomes in sufficient detail to be useful. Data is collected and compared to these objectives. The discrepancy analysis is then performed (what should be (-) what is, (=) equals the discrepancy). Discrepancy information might lead to a change in the operation of the program or in the specification under which the program operates; or to better control the processes in the training environment; or be used to re-design the system and/or command's training baseline and process, and/or its relationship to the military service's training specification.

Training needs assessment data is initially derived from engineering documentation, engineering requirements and system procedures. Feedback is solicited from the command concerning engineering changes, procedural changes and technician operational problems; and from knowledgeable personnel affiliated with the system hardware, equipment or personnel operations, and technician supervisors. This often occurs in regularly scheduled evaluation committees as in the case of the Navy's FBM and AEGIS programs, composed of engineering, management and training personnel. The evaluation process also includes operational performance records which are analyzed and compared to baseline requirements. Data is also obtained from the training program records concerning trainee performance within courses. The model incorporates a training effectiveness algorithm to detect man-machine interface problems within each of the training program components. The model also employs a task evaluation algorithm to assess task relevance within courses in the program.

This Discrepancy-Based Model is three-phased. The First Phase permits program evaluators to systematically review and pinpoint lack of congruence in the command and/or system's program baseline with accepted industry training and military service standards. Phase Two of the model assesses individual components of the training program in order to identify program congruence to stipulated baselines. Phase Three then allows for evaluation of training program outcomes, and description and discussion of the findings.

The Model:

Phase One - A System Baseline Standard:

Military training is developed to specific baseline requirements and standards. The documents which constitute baseline training standards define the organizational relationships and authority between the command and its' training activity. The baseline includes a training program management plan. Policy and procedures are stipulated for training management decision

making, training organization composition, and inter-organizational relationships. The baseline provides a process for needs assessment including job and task analysis. These components serve as a foundation for training development and evaluation. A major part of this baseline is the maintenance of records both as an evaluative database source and as a legal requirement.

The first evaluation is of the command and/or system's baseline standard for management and operation against industry accepted standards and the particular service's procedures. To assess the command's baseline standard, Phase One includes a systematic multiple data source discrepancy analysis of degree of implementation of each of the program components against the system and/or command's program documentation. This process is best conducted by the review team prior to arriving on the scene at the command. Individual and independent review by each member is aimed at capitalizing on the individual team member's expertise to arrive at appropriate conclusions without external influence from the other members. This data is treated in a very rigorous and discrete manner to ensure that only appropriate and indicated changes are suggested for the training program. The team next assembles at the training site and begins review. The objectives of this is to come to a group consensus on the first phase findings of the baseline training standard. Depending on whether this is a routine review or an external command review, the report generated will contain recommendations for formative changes, or a summative report on the state of the baseline training standard.

Phase One of the Discrepancy-Based Total Training Program Evaluation verifies the existence of baseline requirements to a discernable level. These criteria will remain a functional baseline over the course of the total evaluation. Discrepancies which are identified in the system's baseline documents will be noted in accordance with the command criteria.

PHASE TWO - Course and Instructional Review:

Phase Two of the process allows review of specific courses, instructional processes, training plans, trainee records, and other components of the training program. To assess the effects of the system's training processes, this evaluation tool must allow review of the engineering and operations data for indications of training related problems, and indications of needs for training to alleviate engineering faults, and then provide prescriptive formative findings of program outcomes. It also must be capable of providing overall summative assessments; indicators of overall program success in order to address reporting requirements to command and military decision makers.

A significant aspect of the formative component is operationalized through the individual course evaluations. The Discrepancy-Based Total Training Program Evaluation incorporates aspects of the Instructional Quality Inventory (Montague, et. al., 1983), a formative evaluation tool developed for US Navy courses. Built upon the principles of instructional systems development, IQI provides an empirical methodology for course

evaluation using pre-identified and stated behavioral objectives. This process permits a comprehensive auditing of courses and programs against the learning requirements for the course(s).

To provide for more comprehensive auditing of the entire training program, including multi-dimensional learner and organizational requirements, training media, and numerous courses, the Training Effectiveness Algorithm (Cantor, 1986) (Figure 1) compliments the IQI and permits the identification of the causes of reported training problems within the training system and ongoing engineering organization. As a two step procedure it uses job incumbents and available resources to analyze and identify training related problems.

The Task Evaluation Algorithm (Cantor, 1985) (Figure 2) was adopted to ensure that the process used to evaluate the methods for curricula modification and incorporation of up-to-date job analysis data is effective and organized. This process permits a systematic coding of all tasks by grouping tasks into departmental requirements, and reviewing tasks against existing lesson plans. This model component provides a means for auditing curricula and lesson plans against specific tasks required on the job.

In Phase Two the evaluator analyzes each of the areas of the system's training program and process against the established baseline. An analysis is performed in each of the following areas:

- Needs Assessment data to determine if the processes are followed, files maintained and key personnel are involved, etc.
- Performance Objective development processes is in place.
- Curricula development, lesson plans, media selection methods, etc. are verified using lesson plan files selected randomly from program files.
- Facilities, classrooms, labs, simulators are all personally observed.
- Instruction is formally observed and evaluated.
- Staff qualifications and development is reviewed and randomly selected staff records are selected and reviewed.
- The training evaluation program is carefully studied and individual data trails followed to determine the means of data utility for program revision.

The outcomes of Phase Two will be written and presented as indications for corrective action (formative findings) or in the case of military service and/or command inspection (summative findings).

PHASE THREE - Analysis & Outcomes Reporting:

Phase Three of the evaluation synthesizes the findings of the phase two reviews and analyses. All data supporting the system analyses are compiled and reviewed. At this point the discrepancies are noted in each of the evaluation categories. Careful attention is given to review of data concerning course

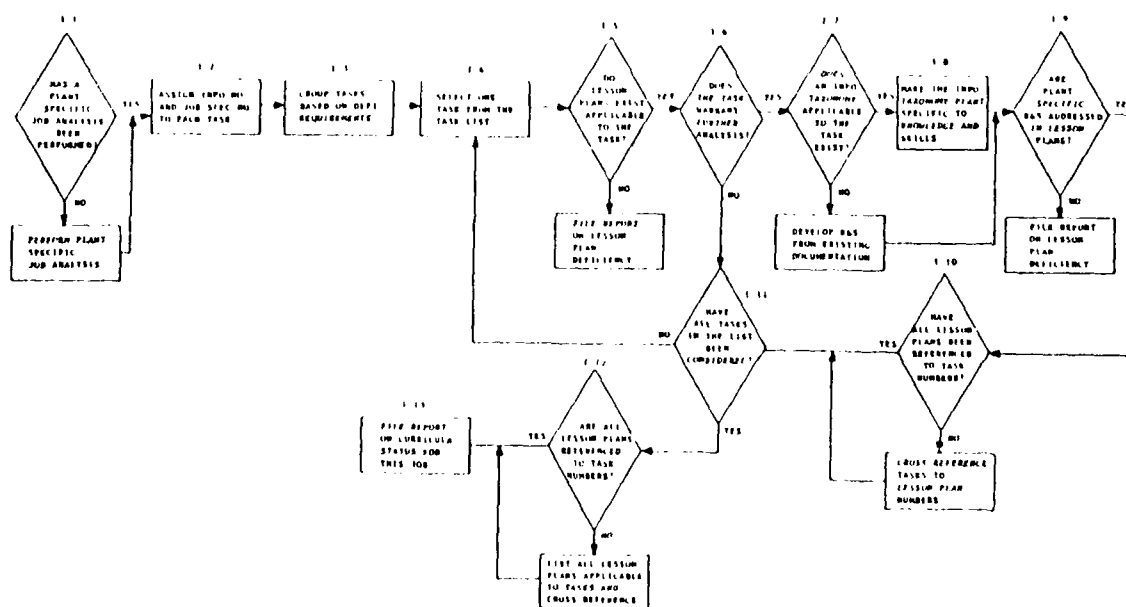
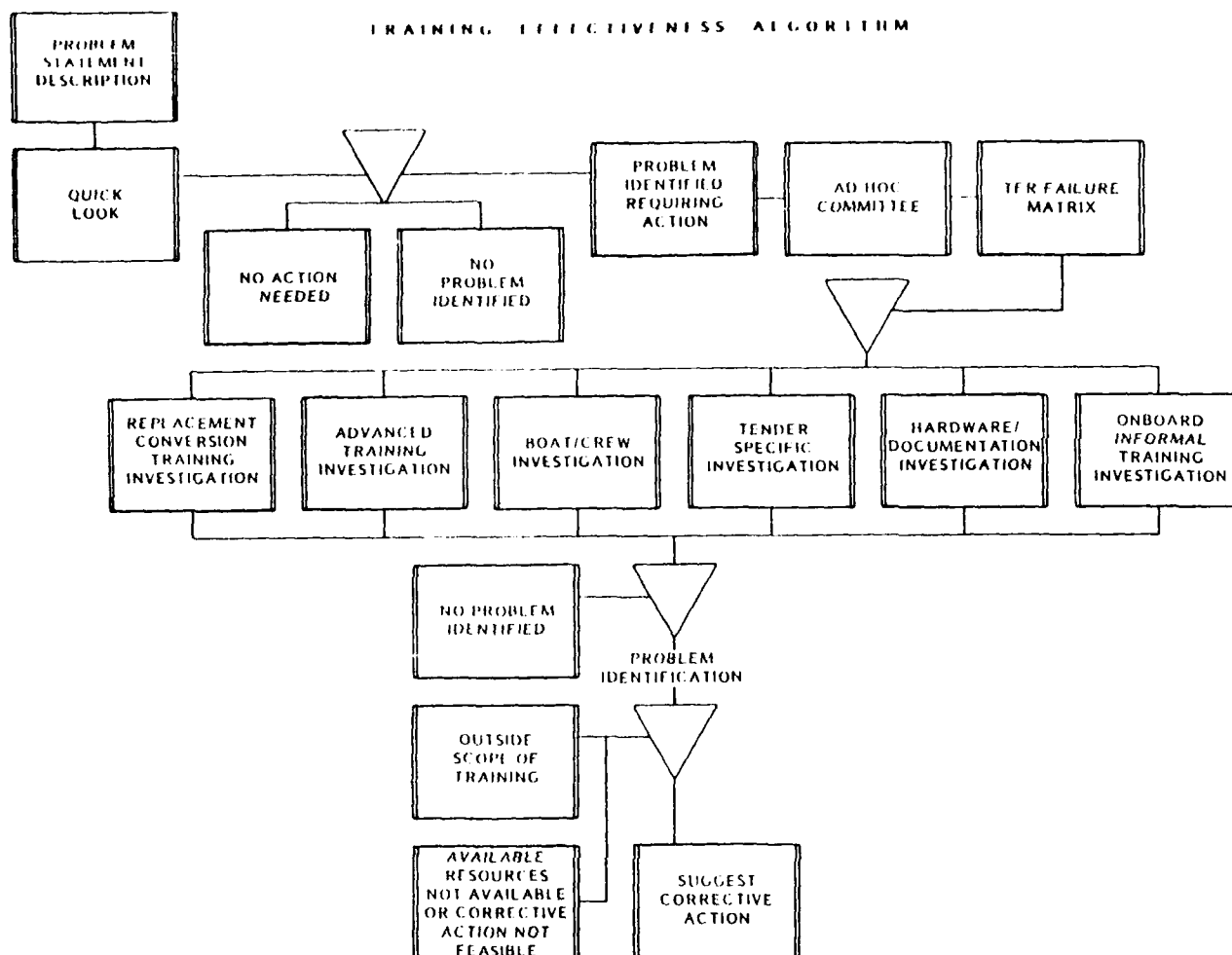


Figure 1. Comparing Existing Curricula to Job Analysis Results.

and program congruence to baseline requirements, and to course and program outcomes' and effectiveness. Data analyses of course and lesson plans are reviewed, lesson observation notes are studied, facilities reviews considered and final conclusions reduced to writing.

The final report is two-fold and includes: (1) report of discrepancies between best commercial practices standards and command standard; and (2) report of discrepancies between command standard and/or system training operations. Emphasis is placed on the areas and problems which training system management should prioritize and incorporate into program revision. In as much as these evaluations often become evidence in deciding the future of systems and programs, as well as court subpoenaed information, very finely detailed findings, discrepancies, and problems are provided. The reporting procedure often involves an extensive briefing to command officials, training management, and sometimes public officials.

Applicability:

The purpose of this paper was to describe, discuss and analyze a Discrepancy-Based Methodology for Total Training Program Evaluation. All too often training program reviews amount to nothing more than cursory notations of individual evaluation perceptions and biases. Substantive data, if available, is used to support preconceptions about programs and their outcomes. The discrepancy-based program evaluation can be viewed as a fresh empirical approach to training evaluation due to its ability to ensure a rigorous and systematic framework for analysis of individual program components against a recognized baseline of measurable program objectives. The use of an expert team approach to the process is another plus. No one voice can become the theme of the findings. The process also allows for both formative corrective findings of specific program components as well as overall summative reporting. In essence this provides a service to engineering and training managers as well as policy-makers.

Bibliography:

Cantor, J.A. (1985) Task evaluation: comparing existing curricula to job analysis results. Journal of Educational Technology Systems, 14 (2), 157-163.

Cantor, J.A. (1988) The training effectiveness algorithm. Journal of Educational Technology Systems, 16 (3), 207-229.

Montague, W.E., Ellis, J.A., & Wulfeck, W.H. (1983). The instructional quality inventory (IQI): A formative evaluation tool for instructional systems development. Monograph: Navy Personnel Research and Development Center, San Diego, CA.

Civilian Test Experts Help Improve Skill Qualification Tests

Paul R. Vaughan and Clay V. Brittain

U. S. Army Training Support Center, Fort Eustis, Virginia

Purpose

This paper describes the efforts by civilian test experts contracted under the Scientific Services Program (SSP) to improve the Army's Skill Qualification Tests (SQTs). The different contributions and general recommendations resulting from the SSP efforts are presented.

Background

Skill Qualification Tests (SQT) are used by the United States Army to test enlisted soldiers in their ability to perform selected tasks of their Military Occupational Speciality (MOS). The use of SQTs began in 1977. The tests are designed to serve two major purposes: (1) to provide scores for use in personnel decisions, including the selection of soldiers for promotion to the next higher rank and for retention in the Army, and (2) to provide feedback useful to commanders in training soldiers. Initially SQTs included hands-on performance testing as well as paper-and-pencil multiple-choice testing. But resource intensiveness and the practical problems of hands-on performance testing led to a decision in the early 1980's to modify the program. The Army adopted what is called the Individual Training Evaluation Program (ITEP) which includes the SQT as one of its three components. The other two components are the Common Task Test (CTT) and Commander's Evaluation. Both of these two components are hands-on evaluations. As a part of the ITEP the SQT has become an all-written test.

Skill Qualification Tests are developed by Army service schools located throughout the United States. There are some 21 schools responsible for developing SQTs. The Army Training Support Center (ATSC) at Fort Eustis, Virginia is the Army's

The views, opinions, and/or findings contained in this report are those of the authors and should not be construed as an official Department of the Army position, policy, or decision.

program manager for ITEP and in this role is responsible for evaluating the SQTs for purposes of quality assurance. These responsibilities are executed by the Individual Training Evaluation Directorate (ITED) of ATSC. To supplement test expertise in the Army, the ITED has used the SSP contract to obtain the services of experts from the civilian sector.

The SSP Program

The Scientific Services Program (SSP) is operated by Battelle Columbus Laboratories (BCL) under contract with the Army Research Office (ARO). The program is described in the ARO brochure "Scientific Services Program" dated March 1987. The SSP contract is designed to provide federal agencies short-term nonpersonal services for the solution of scientific and technological problems.

Once a requirement is established for services under the SSP, the agency must prepare a request for submission to the ARO Contracting Officer who, after acceptance, transmits the request to the prime contractor, BCL. The prime contractor is responsible for performance of all work, including selecting and negotiating with various subcontractors to perform the effort.

The requesting agency must transfer funds to ARO to cover costs of the services requested plus an 11 percent overhead for the prime contractor. An individual designated by the requesting agency is appointed as the Contracting Officer's Representative (COR). The COR monitors the work of the subcontractors, whose claim vouchers must be certified by the COR prior to payment.

What We Have Asked from SSP

Beginning in 1982, experts from the civilian test community have been used to conduct psychometric evaluation of the SQT program. Additionally, they have visited schools to meet SQT developers to give them face to face assistance in test improvement.

We have asked the test experts to: determine judgmentally the extent to which the task tests cover the task elements; review each SQT question to determine if it conforms to sound psychometric principles and good testing practice; provide specific recommendations for correcting deficiencies and; give an overall evaluation of each SQT selected for review. Also, we have called upon the SSP contractors to recommend any modifications of procedures and approaches to SQT development which seem likely to improve the quality of SQTs.

Extent of Contributions by SSP Reviewers

The evaluations generally have been thorough and the evaluators have offered concrete suggestions for improving the test items. However, this has not been true across the board. In several instances, the reviewer's comments and suggestions did not go beyond what could be rather easily inferred from the test item statistics provided to the reviewer. This variation in quality of the evaluations may be related in part to the technical nature of the MOS. There seems little doubt that lack of technical knowledge is a severe handicap in evaluating highly technical SQTs. Many of the evaluators have commented on this point. Nonetheless, many evaluators have been able to comment insightfully on technical SQTs.

The SSP evaluators have contributed a great deal in three areas. First, they have instructed item writers in the common rules for writing good multiple choice items. Second, they have helped the SQT developers to more effectively use good principles of testing and item statistics in constructive criticism of a test. Third, they have suggested modifications to the Army policy on SQT development.

Recurring Themes from SSP Reviews

Early on, several recurring themes emerged from the SSP evaluations of SQTs. Since the reviewers saw their central role as that of finding faults, the recurring themes tended to be negative.

The evaluators' critical comments about SQT and their recommendations for test improvement fall generally into two categories: those having implications for SQT development and those having implications for SQT policy.

Problems Related to SQT Development. Major points noted by the SSP reviewers are as follows:

Construct Validity - In many instances the test may be measuring only the soldier's ability to read and find the correct answer in references provided.

Content Validity - Task coverage is uneven and in some instances is inadequate. The task tests are often too short to provide a reliable decision about the soldiers ability to perform the task (Hambleton, 1984).

Reading Load - The reading burden on some SQT is excessive.

Distribution of Correct Answers - Keyed answers are more frequently option B or C. This favors the test-wise soldier.

Punctuation, Spelling, Format - Inconsistencies within an SQT are bothersome.

Phrasing of Answer Options - Must be clear, logically consistent, readable and non-cueing. Many decoy answer choices are implausible making selection of the correct answer too easy.

Problems Related to SQT Policy. The following points relate to SQT policy:

Varying number of options per question - Many evaluators felt that mixing of two-, three-, four-, and five-option items in the SQT increased clerical errors and recommended against it.

Extensive overlap of SQT across skill levels - The SSP experts noted that for some MOS, SQT 2 is almost identical to SQT 1 and SQT 4 almost the same as SQT 3. The reviewers pointed out that the higher skill levels performed different jobs, therefore, the tests used to evaluate them should be more unique.

Building item banks - Evaluators recommended the systematic development of computer-based item files for production of SQT from year to year. This will improve tracking, identification of problem items, and allow better documentation.

Improving item analysis - The item analysis is inadequate and there is a lack of training to interpret the information. There is strong need for additional statistical information on reliability and item discrimination.

Limitation of multiple-choice tests - Need to continue exploring testing options to supplement multiple-choice testing.

The above points relating to policy are of particular interest to the ATSC managers of the SQT program. With regard to both test development and policy the reviewers did not always agree on the best ways to improve SQT quality. For example, some experts suggested that tasks be randomly sampled from the job domain while others wrote that using expert judgement would improve content validity. Also, they were divided on the issue of whether varying the number of options in the same test resulted in significant clerical error.

SQT STRENGTHS. Even though the emphasis was on negative aspects of the tests, the reviewers did notice some strengths. The following comments are illustrative:

In many cases, the SQT reviewed did an excellent job of making the tests job-specific and similar to "real world on-the-job" requirements.

The SQTs reviewed are clearly performance-oriented, appear to have adequately sampled the critical task domain, and have followed generally accepted conventions of item construction and format.

Use of SSP Reports by Schools

Individuals involved in SQT development vary on their reactions to assistance rendered by the SSP reviewers. Some seem to appreciate the recommendations very much, while others state that the suggestions are not useful since their school's education specialists and other personnel can make most of the same recommendations. Generally, the negative responses have disappeared in the last two years.

The ITED conducted a study in an attempt to better understand the usefulness of the civilian test expert's recommendations. We used SSP evaluation reports containing suggestions for test improvement to discover to what extent recommendations were followed by the developers during the next iterations.

More than half (54 percent) of the items targeted for improvement by the SSP reviewers were changed as recommended. One school followed every recommendation, and one school used only 15 percent of the item change suggestions. About 50 percent of the suggested changes were adopted by each of the other eight schools involved in the study.

Hills (1987) suggested possible reasons to account for the lack of change from SSP suggestions over the years:

1. Lack of adequate training in test development, often a result of personnel turbulence.
2. School may not accept civilian suggestions as being related to Army problems.
3. School developers may be defensive about "their" test items.
4. Suggestions may be too expensive for resources available.
5. Individuals in charge of the test program may not support the recommendations.

Resolutions & Conclusions

The Army SQT developers are now actively pursuing most of the SSP recommendations regarding SQT development. The test reading loads are much more appropriate now than earlier and generally the distribution of correct answer choices has been even. Editing of the SQTs has improved considerably. However, the program is still troubled by too many implausible distractors that make the questions too easy.

There are also efforts underway to implement several of the changes to policy recommended by the evaluators. Recently the number of answer options has been more constant. The questions usually have four answer choices. Item banking software is being field tested by several schools with a goal of having automated item banks in each of the schools by 1990. The item analysis produced by the ATSC has been vastly expanded to improve usability as suggested by the SSP evaluators. It appears that the use of outside assistance has been successful in improving SQT quality.

References

Hambleton, R.K. (1984) Determining test length. Berk, R.A. (1984) A guide to criterion-referenced test construction. Baltimore: John Hopkins University Press.

Hills, J.R. (1987, December) Some things that need to be done to make SQTs do what they are supposed to do. Keynote speech presented at the U.S. Army Training Support Center annual ITEP conference, Fort Eustis, VA

THE U.S. MARINE CORPS INFANTRYMAN JOB PERFORMANCE MEASUREMENT PROJECT

Milton H. Maier, Chair
Center for Naval Analyses
Alexandria, Virginia

The Infantryman Project is part of the U.S. Marine Corps' contribution to the Joint-Service Job Performance Measurement (JPM) Project. The primary purpose of the study was to validate the Armed Services Vocational Aptitude Battery (ASVAB) and a battery of experimental predictor tests against the best possible measures of job performance for five Marine Corps infantry specialties (Infantry Rifleman, Machine Gunner, Mortarman, DRAGON, and Infantry Staff NCO). A secondary purpose was to evaluate the equivalence of various potential surrogate measures of job performance (e.g., paper-and-pencil job knowledge tests, training grades, performance ratings, performance indicators contained in administrative and personnel files) with benchmark hands-on job performance tests. Throughout the project, careful attention was devoted to hands-on testing to ensure accurate, high fidelity measures of benchmark performance. Procedures were developed for selecting tasks for testing, for analyzing tasks and developing hands-on measures, and for ensuring high quality data collection. The results of the project indicate that this diligence paid off. High levels of test-retest and interrater reliability were attained for the hands-on tests. Validity analyses indicated extremely high correlations between the ASVAB and hands-on performance. The papers in this panel describe the design of the project, the procedures used to select tasks for testing and to develop tests, the quality control procedures used to ensure high quality test data, and the results of the validity study.

Task Selection and Test Development* for the Infantryman JPM Project

Daniel B. Felker
Charles W. Harnest

American Institutes for Research

This report relates to a research project conducted by the American Institutes for Research (AIR) for the Center for Naval Analyses, "Developing Job Performance Tests for the United States Marine Corps Infantry Occupational Field." The project was conducted over a 27-month period, beginning July, 1986. During the course of this effort, approximately 150 hands-on and job-knowledge tests were developed and administered to some 2600 active-duty Marines in five Infantry MOS. About half of the testing took place at Camp Lejeune, NC, and half at Camp Pendleton, CA. Additional Measures of performance -- supervisor ratings and training school grades -- and various predictor measures -- notably the Armed Services Vocational Aptitude Battery (ASVAB) -- were also administered. This paper focuses on the selection of tasks to test and the development of hands-on and written job knowledge tests.

The project's goal was to develop measures of USMC Infantry job performance that would serve as criteria for validating predictors. Five Infantry Military Occupational Specialties (MOS) were of interest: 0311 Rifleman; 0331 Machine Gunner; 0341 Mortar Man; 0351 Assaultman; and 0369 Infantry Unit Leader. Tasks designated as 0300, which are common to all Infantry MOS, also were included in the study.

Specification of Job Performance Domains

The initial job analysis thrust was to define boundaries of the USMC Infantry jobs by specifying the performance requirements of each MOS. Performance domains include all tasks that job incumbents in each Infantry MOS are expected to perform.

We relied on two documents to create the performance domains: Individual Training Standards (ITS) for the Infantry Occupational Field and the Essential Subject Manual (ESM). The ITS lists the tasks that Marines in each Infantry MOS are expected to be able to perform, plus information about where

* This research was funded by Contract No. N00014-83-C-0725 and by Subcontract CNA 1-86. All statements expressed in this paper are those of the authors and do not necessarily reflect the official views or policies of the Department of the Navy or the U.S. Marine Corps.

these tasks are trained and the rank when task proficiency is expected. The ITS also clusters tasks within each MOS into Duty Areas, which the ITS defines as "groupings of tasks into units that are typically learned, practiced, and performed together." To illustrate, one Duty Area of the 0300 MOS is Hand Grenades which is composed of three tasks. In addition, the ITS refers to a number of tasks and Duty Areas listed in the Essential Subject Manual (ESM), which lists tasks Marines in any MOS are expected to perform. The performance domain constructed consisted of 183 infantry tasks as summarized in Table 1.

Table 1. USMC Infantry Performance Domain

<u>MOS</u>	<u>No. of Duty Areas</u>	<u>No. of Tasks</u>
0300	15	110
0311	2	10
0331	1	10
0341	2	2
0351	2	9
0369	<u>4</u>	<u>12</u>
Total	26	183

The task descriptions listed in the ITS and ESM varied greatly in descriptive specificity and scope. Using USMC documents, we reworked these task descriptions to produce more consistent and comparable levels of description. This process increased the number of original activities considered as "tasks" for this project.

Identification of Behavioral Elements

Because it was not feasible to test all tasks in the performance domain, we sought ways to stratify tasks in terms of content to help select tasks for testing. USMC Subject Matter Experts (SMEs) further analyzed the tasks to identify underlying skills, knowledge, and procedures required for task performance so that this information could be used for stratifying tasks. We called these task components "behavioral elements," which operationally are verb-noun statements denoting behaviors that underlie performance of tasks. Behavioral elements found in different tasks identify commonality across tasks and lead to predictions that performance on the different tasks will be correlated.

The final job analysis step consisted of developing task-by-behavioral-element matrices for the Duty Areas of each MOS. Tasks made up the column headings of each matrix and behavioral elements associated with the tasks formed the matrix rows. A matrix cell entry indicated that the behavioral element listed for the row occurred in the task listed as the column head.

The matrices illustrated the overlap of behavioral elements within Duty Areas and among the tasks in the entire domain. Conceptually, elements repeated most frequently within a matrix are considered most representative of the tasks in that matrix. Information about overlapping behavioral elements was used for planning task selection procedures.

SMEs from the 2nd Marine Division at Camp Lejeune and the 1st Marine Division at Camp Pendleton reviewed the task-by-behavioral-element matrices for accuracy and completeness. In preparation for task selection, SMEs also ranked tasks for representativeness to Duty Areas and rated each of the 15 0300 Duty Areas for relative importance.

Task Selection

Our goal was to select a set of tasks to test which would generalize to the largest number of tasks in the Infantry. The selection strategy had to accommodate the six hour restriction for hands-on testing of 0300 content and one hour testing limit of 0311, 0331, 0341, 0351, and 0369 content.

The selection strategy for 0311/31/41/51 content was straightforward. After tasks deemed infeasible to test by SMEs due to cost or safety factors were eliminated, the allotted testing time was sufficient to test all remaining tasks of these MOS. For 0369, we selected tasks that could be tested meaningfully by simulation on a war-game board (TACWAR), since this MOS involved supervisory and tactical tasks that are hard to standardize for individual hands-on testing.

For 0300 there was much more content than could be tested in the six hours available for testing. We chose a random sampling strategy built around the following selection rules:

- o All 0300 Duty Areas were to be tested.
- o Target testing time was 360 minutes (6 hours); 250 minutes for actual testing, 110 minutes for administration and logistics.
- o Ceiling testing and test administration times per Duty Area were set as:

- 30 minutes for Duty Areas SMEs ranked most important
- 20 minutes for Duty Areas ranked next most important
- 15 minutes for Duty Areas ranked third most important
- 10 minutes for Duty Areas ranked fourth most important

- o Sampling unit within each Duty Area was the task as defined previously

Our sampling strategy maximized test content coverage across the behavioral elements, while not excluding the possible replication of behavioral elements. Emphasis thus was placed on measurement of the underlying behaviors associated with job performance as well as with specific tasks. This contrasts with the customary practice of focusing only on the task dimension when selecting performance test content.

We followed a meticulous sampling process. Each task in a given matrix was assigned a weight equal to the sum of its behavioral elements. Task titles were printed on a number of paper slips equivalent to that task's weight. Thus, for a task with 16 behavioral elements, 16 slips of paper with that task's title were prepared and thrown in a pool with the slips of paper for all other weighted matrix tasks.

A task was "sampled" by drawing a slip from the pool. At that point, all remaining slips with that same task title were removed from the pool. We then examined the behavioral elements that formed the sampled task and crossed those behavioral elements off the matrix. The remaining tasks were reweighted by the number of remaining behavioral elements and the equivalent numbers of slips of paper were reentered into the task pool. After each drawing, the estimated time to test the sampled task was subtracted from the total remaining allotted time for the Duty Area. Tasks were sampled and the process repeated until testing time was exhausted. This entire procedure was repeated for the parallel test form.

The percentage of behavioral elements selected out of the total number of possible behavioral elements for each 0300 Duty Area ranged from 100 percent to 42 percent for one test form and from 100 percent to 47 percent for the alternate form.

Development of Tests

Task selection established the general content area to be tested; it did not indicate what would be scored as pass or fail on the hands-on test. We conducted task analysis to

determine the steps necessary to complete each task, the sequence of step performance, and the equipment and materials needed for task performance. SMEs reviewed the results of the task analyses for completeness and accuracy.

Development of hands-on tests proceeded by converting task analysis information from each task into a "testable unit" (TU). All TUs consisted of a Set-up Sheet and a Scoresheet. The Set-up Sheet provided information needed to conduct the test; it listed the equipment and materials needed, instructions for setting up the test, instructions to follow after each examinee completed a testing station, and special instructions for administering and/or scoring the test. The Scoresheet was used for recording examinee identifying information and performance scores. It included two columns for scoring examinee performance on each step as "GO" or "NO GO."

A total of 146 Testable Units was developed spread across MOS and Test Forms as shown in Table 2. TUs were pilot tested with different groups of five infantrymen over a three month period at Camp Lejeune. TUs then were refined and prepared for the field research phase of the project.

Table 2. Number of Testable Units Developed

MOS	Test Form			
	AB	A	B	
0300	40	25	21	
0311	7	0	0	
0331	11	0	0	
0341	13	0	0	
0351	13	0	0	
0369	<u>0</u>	<u>8</u>	<u>8</u>	
Total	84	33	29	= <u>146</u>

Paper and pencil job knowledge tests (JKTs) were developed to parallel hands-on test content as much as possible. The purpose for equating content on the two test modes was to provide a basis for determining whether paper-and-pencil measures are sufficiently related to the more costly and time consuming hands-on measures to justify using them as surrogates.

Development of JKTs started with the performance steps listed in the hands-on tests. Test items were written around specific steps or groups of steps critical to task performance.

The items were mostly written as multiple-choice questions, but other formats were used when appropriate. Because of the need to create surrogates for hands-on tests, emphasis was placed on developing written test items that were performance-based. Test items stressed what and how steps are performed rather than why. Extensive use was made of graphic materials and illustrations to maximize the performance dimension. For tactical tasks and other nonprocedural tasks, we developed combat scenarios which asked questions about what should be done on the basis of information given in the scenarios.

Two test forms were developed corresponding to the 0300 content for Form A and B hands-on tests. One test form only was developed for the 0311, 0331, 0341, 0351, and 0369 MOS.

The JKTs were pilot tested on 71 Marines at Camp Lejeune. Item analyses were conducted to identify items to revise or delete. Items were dropped if they had: 1) one or more distractors with a high positive biserial; 2) low pass rate (low percentage endorsement of correct option; and 3) negative biserial (Brogden-Clemans biserial).

The final tests for 0300 consisted of 150 items which were expected to be completed in 90 minutes. The number of items for the 0311, 0331, 0341, and 0351 MOS-specific job knowledge tests range from 40 to 50 and were expected to be completed in 30 minutes. The 0369 job knowledge test contained 142 items and were to be completed in 90 minutes.

The hands-on and job knowledge tests were put in final form, printed, and distributed for formal tryouts and testing over four months at Camp Lejeune and Camp Pendleton.

Quality Control Procedures and Interrater Reliability Results*

Jennifer L. Crafts
Edmund C. Bowler
David W. Rivkin

American Institutes for Research

The need for quality control procedures in the conduct of hands-on testing was highlighted in the U.S. Marine Corps' (USMC) initial validation research effort. Maier (1988) reported that in the study of two MOS (radio repairers and automotive mechanics) where the Marine Corps had responsibility for providing the examinees, facilities, and test administrators, the obtained validity coefficients were reduced by the error-filled data. Various measures (e.g., checks for outlier scores and missing scores, deletions of inappropriate subgroups, corrections for differences in test administrator scoring standards, etc.) were used during the analyses to correct for these low initial validity coefficients. Maier (1988) concluded that quality control procedures should be applied when collecting data in a field setting. In the evaluation of the Joint-Service Job Performance Measurement/Enlistment Standards Project, the Committee on the Performance of Military Personnel also emphasized the need for quality control measures (Wigdor & Green, 1986). The committee viewed testing at a variety of sites, noting conditions that threatened the quality of data collected. The committee then suggested some guidelines for improving testing conditions and selecting and training test administrators.

The USMC Infantryman Project conducted by AIR built procedures into the design so as to circumvent as many of the problems encountered during the initial study as possible. As we will illustrate later in this presentation, the quality control measures were very successful. Both reliability and validity results were commendable.

We will describe the quality control measures instituted to ensure the collection of high quality data. These measures included recruiting retired Marines to serve as Test Administrators (TAs), conducting structured interviews for these TA positions, conducting extensive hands-on TA training, rotating TAs among testing stations, shadow scoring examinee performance, monitoring TA performance (with feedback), and maintaining TA morale. We will then summarize results of analyses of interrater reliability.

Recruiting of Retired Marines

We recruited retired SNCOs and officers from nearby the two testing locations to serve as test administrators. We based our decision to hire retired Marines on previous experience in military testing. Most previous

* This research was funded by Contract No. N00014-83-C-0725 and by Subcontract CNA 1-86. All statements expressed in this paper are those of the authors and do not necessarily reflect the official views or policies of the Department of the Navy or the U.S. Marine Corps.

projects have used incumbent NCOs or civilian researchers as TAs. Both of these groups have obvious weaknesses. We anticipated that retired USMC SNCOs would be superior because of their motivation, knowledge of Marine Corps subject matter and procedures, dedication to the project, and reliable attendance. Through both formal and informal Marine Corps networks, we recruited approximately 120 interested candidates and set up interviews with 116 of them at the two sites.

Structured Interview Format for Test Administrator Positions

A structured interview format was used by three interviewers at one testing site and two interviewers at the second site. Each 30-minute (approximately) interview began with the interviewer giving the applicant a brief description of the project and the indoor and outdoor TA positions. Each applicant was asked the same series of questions to obtain responses about the applicant's experience in a variety of areas critical for the test administrator job, e.g., familiarity with test administration procedures, administrative skills, conscientiousness and compliance, adaptability and initiative, public speaking/reading ability (applicants read samples of hands-on TA instructions), infantry experience, and familiarity with new infantry weapons systems. For these first six dimensions, interviewers assigned a rating from 1 (low) to 5 (high) to represent the applicant's degree of experience on those dimensions. For the dimension "Familiarity with new infantry weapons systems," a mean rating was derived to reflect the applicant's familiarity with nine new weapon systems, each measured by a three-point scale (1 = Not at all familiar, 2 = Moderately familiar, 3 = Extremely familiar).

Mean ratings on these dimensions were used to classify applicants as: (1) reject, (2) possibly hire, and (3) definitely hire. Offers were extended to the top applicants at each site; 34 and 32 accepted positions at Camps Lejeune and Pendleton (respectively).

Test Administrator Training

We invested extensive time and resources in TA training. The training program was designed so that TAs (a) learned basic concepts of test administration, (b) received extensive behavioral practice in actually performing the tasks, and administering and scoring the tests, and (c) received specific, immediate feedback on performance during the training sessions.

Training was conducted over five days in a realistic setting. Equipment that was actually to be used in the testing was obtained for training. Emphasis was placed on TAs becoming totally familiar with all steps of all tests they were to administer. "Indoor" hands-on TAs practiced scoring for all tests scheduled to be tested inside, while "Outdoor" hands-on TAs practiced scoring all tests scheduled for testing outside.

Throughout the training week, trainers emphasized the importance of standardized test administration procedures, objective scoring procedures, withholding feedback from examinees during and after testing, completing scoresheets correctly, maintaining test security, and maintaining a professional manner while administering tests.

Another important goal of training was to conduct, to the greatest possible degree, identical training at the two different test sites at two different times with two sets of TAs. To ensure training equivalency, the same AIR staff conducted training at both sites, utilizing the same training strategy. TAs were exposed to the same instructional techniques and procedures. Most importantly, TAs learned and practiced test administration under the same scoring standards.

The training for each "test package" (group of tests administered at the same test station) consisted of several steps. First, the trainer covered equipment requirements, instructions for setting up the station for the start of each test administration, and administration and scoring procedures. The trainer then demonstrated the task, explaining each detailed performance step on the scoresheet. TAs took turns scoring, administering, and performing the test. While one TA took the examinee role and a second served as the administrator, all other TAs acted as "shadow scorers." Shadow scorers scored all steps while the "primary" TA administered and scored the test. At the completion of each "test," the trainer and all TAs compared their scoresheets and critiqued TA performance. Inter-scorer differences in interpretation were discovered and reconciled by discussing judgments on a step-by-step basis. The rotating and scoring process was repeated until all TAs were comfortable with the test as well as proficient at administration and scoring procedures.

Shadow Scoring

Perhaps the most important quality control procedure we implemented was shadow scoring: scoring performance of the same Marine performing the same test by two scorers. Shadow scoring was systematically conducted during the tryout and field test in order to monitor TA performance and test reliability. Each day, one outdoor HO TA and two different TAs from the indoor HO group were designated "shadows." The shadows randomly selected an examinee at the start of a testing session to follow through all stations. Shadows silently scored performance while the primary TA assigned to each station administered and scored the test.

Test Administrator Rotation

In order to reduce boredom and maintain motivation levels over the four-month data collection effort and to reduce any systematic effects of TA individual differences, we rotated TAs through stations. Every one-half to two days, TAs rotated through their respective stations. TAs thus administered tests at all applicable stations and functioned as shadow scorers an approximately equal number of times.

Test Administrator Performance: Monitoring and Feedback

Observation of TA Performance. A dress rehearsal of all testing procedures was held after training was completed at each site. Thus, TAs were able to practice all procedures under realistic conditions. Training staff observed TA performance and gave feedback regarding coaching, etc., as necessary. Managers of the various testing segments continued to monitor TA performance throughout the field test.

On-site data entry. AIR developed a computer-based Hands-On Score Entry System (HOSSES) to enter, verify, and report analyses of collected data. Daily estimates of TA consistency were computed. Entering data on-site rather than after completion of the testing allowed us to (a) monitor the data for completeness; (b) identify any problem tests or steps; (c) immediately follow-up on and provide solutions to identified discrepancies, inconsistencies, and missing data with the TAs; and (d) use the reports of TA consistency to give performance feedback to TAs both individually and as a group.

Feedback from Managers. The HOSSES program automatically produced three reports after data were entered and verified. These included:

1. DETAILED DISCREPANCY REPORT - a listing of all steps where the shadow and primary scorers disagreed.
2. DISCREPANCY SUMMARY REPORT - a listing of the percent agreement with the shadow scorer at each station for each test administrator.
3. SUMMARY BY TASK AND BY TEST ADMINISTRATOR - a listing of the percentage of "GO" scores for each TA on each test.

Given these reports, the Test Site Manager (TSM) reviewed results and identified problem stations, tests, steps within tests, and/or scorers. The TSM or Hands-on Manager (HOM) discussed problem areas with the TAs. The HOM would monitor further scoring to ensure that TAs were interpreting test performance according to the test specifications.

TA Morale Boosters

A number of measures were instituted to keep the level of motivation of the TAs high throughout the field test period. A Station Master (TA) was assigned to each test station and was responsible for setting up the assigned station, collecting equipment, supplementary supplies, and preparing score-sheets for the day. The Station Master concept gave TAs a sense of purpose. Additional non-monetary bonuses provided for TAs were well received. Over the months of testing, these included caps with the USMC logo printed on them, jackets, and free coffee and soft drinks.

Reliability Results for Hands-On Testing: Interrater Reliability

We calculated interrater reliabilities--scorer agreement--to check on the accuracy of TA scoring. Two general questions directed the investigation of scorer agreement. First, we wanted to know whether there were content effects: were there differences in scorer reliability over the different Duty Areas? The second question concerns temporal effects: did scorer agreement vary across time?

The scored step--TA observation of step performance--was the unit of analysis. The data base included all observations for which both primary and shadow scorer data were available. If either scorer failed to score a step, that observation was not used in computing the percent agreement. We used percent agreement as the index of scorer agreement. Percent agreement was computed for each duty area, using the following formula:

$$n_a$$

$$P_a = \frac{n_a}{n_a + n_d}$$

where: P_a = the percent agreement
 n_a = number of steps which were scored the same by the primary and shadow scorer, summed across all examinees and across all tests administered for the Duty Area
 n_d = number of steps which were scored differently by primary and shadow scorer, summed across all examinees and across all tests administered for the Duty Area

To calculate agreement percentages, the test packages administered at each station were decomposed into tasks, so that all task observations for each duty area could be summed across all examinees. Comparisons of relative reliability between Duty Areas cannot be made because the examinees and scorers varied among the tests composing the Duty Areas.

The graph of scorer agreement by Duty Area for 0311 Infantry Rifleman (Figure 1) shows that reliabilities ranged from 85 to 99 percent for both bases. These are acceptable reliability levels, showing that TAs were able to differentiate between "Go" and "No Go" performance very reliably. These levels are comparable to those obtained for hands-on testing by other branches of the military. The Navy reported agreement between two-person rater teams for hands-on performance tests for two jobs--Machinist Mate and Radioman. Reliabilities, calculated as correlations between rater pairs' scores, ranged from .95 to 1.0 for machinist mate ($n = 108$) and .85 to .99 for a pilot study with radiomen (n unavailable). Although there was some variation among duty areas, the percent agreement was very similar across duty areas and test sites. Similar high reliabilities were found for the other four MOS. The same general scorer agreement pattern was found for the other MOS: (a) minor variation between sites, with one site (Pendleton) having slightly higher scorer agreement; (b) reliabilities ranging from 85 percent to near 100 percent; and (c) minor variation among duty areas.

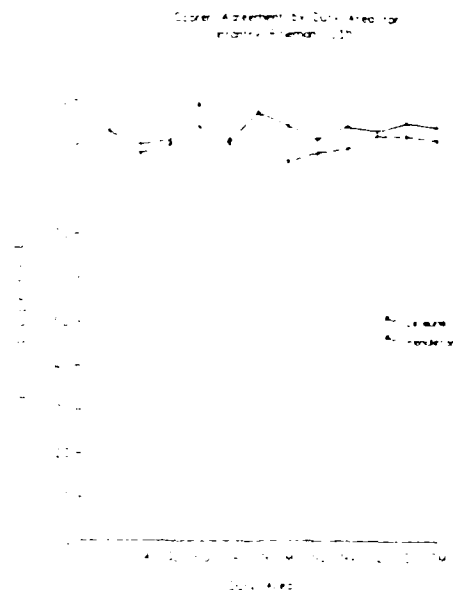


Figure 1. Scorer Agreement by Duty Area for Infantry Rifleman (0311).

Scorer agreement across time was investigated by averaging Duty Area agreement percentages within each test session to obtain a summary index for each session. Scorer agreement over time is shown in Figure 2. This figure shows change over time in the order in which the MOS were tested. The majority of session agreement percentages were in the 86 percent to 97 percent range, with a few low values (81%, 82%) at one site. There was no discernible apparent trend, e.g., increased or decreased rater agreement, over time. We had expected to find a gradual increase over time in percent agreement as scoring problems were identified and TAs received remedial training.

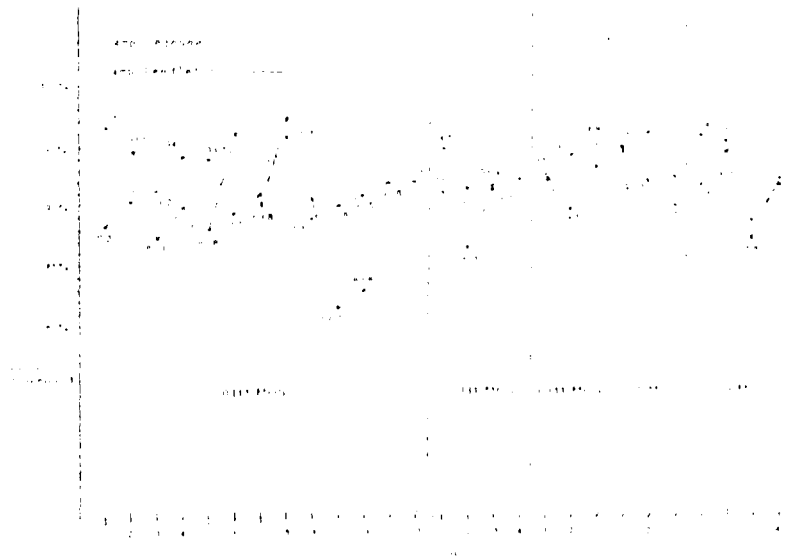


Figure 2. Percent Agreement Across Test Sessions.

Conclusions

The likely explanation for the consistent scoring agreement was that our initial training provided all the instruction and behavioral practice necessary to obtain such high levels of agreement. We also feel that on-site monitoring of TA performance maintained TA vigilance and motivation. In addition, the tryout enabled us to iron out any potential major problems, and scoring discrepancies were investigated and corrected when they cropped up. The 90-95 percent agreement range may represent a realistic ceiling for hands-on tests, with no potential for increase by applying additional motivators or interventions.

References

- Wigdor, A. K. & Green, B. F. (Eds.). (1986). Assessing the performance of enlisted personnel: Evaluation of a joint-service research project. Washington, DC: National Academy Press.
- Maier, M. H. (1988). On the need for quality control in validation research. Personnel Psychology, 41, 497-502.

Quality Control Procedures and Reliability Results*

Andrew M. Rose
Jennifer L. Crafts

American Institutes for Research

There were two principal tenets guiding the conduct of the U.S. Marine Corps Job Performance Measurement Project: scientific rigor and quality control. During all phases -- job and task analyses, task selection, test development, and test administration -- we tried to incorporate techniques and procedures that would lead to the best possible measures. This paper will focus upon several of these techniques and procedures affecting the quality of the resulting job performance measures.

Quality Control Procedures

Experimental Design Controls

During the test administration phase, each Marine went through four half-day sessions. Two consisted of hands-on tests; one involved indoor tests, the other outdoor tests. The other two involved several written tests, including Job Knowledge tests, the ASVAB, and other written predictor tests. We constructed an experimental design that called for one hands-on and one written session each day. The design balanced the order of hands-on and Job Knowledge written test sessions, so that an equal number of Marines received a hands-on and written session first. Also, an equal number of Marines received the indoor and outdoor tests first; likewise, an equal number received the ASVAB and Job Knowledge tests first.

This design controlled for various factors that could affect the validity and reliability of the results, including session order (especially the effect of Job Knowledge and hands-on test order), time of day, and fatigue.

Another counterbalancing design was implemented to control for the possible effects of specific test (as opposed to session) order. During a four-hour hands-on session, each Marine was tested at seven stations. Rather than test Marines in a constant order (e.g., Station 2 always following Station 1), we developed a balanced set of station orders. This set had the property that, over the entire test administration phase, all pairwise orders of stations occurred equally often (e.g., Station 1 followed Station 2 as often as Station 2 followed Station 1).

* This research was funded by Contract No. N00014-83-C-0725 and by Subcontract CNA 1-86. All statements expressed in this paper are those of the authors and do not necessarily reflect the official views or policies of the Department of the Navy or the U.S. Marine Corps.

Examinee Motivation

One of the most difficult aspects of conducting military field research is to properly motivate examinees to perform at the level of their abilities. This is particularly challenging when examinees are told that test performance will not affect their careers and that test results will not be shown to their superiors. We have found that the best way to keep examinee motivation high is to have an officer brief them at the start of the exercise. During this project, each group of examinees was briefed by an officer when they arrived for testing. In addition to normal logistic considerations, this briefing addressed four main points relating to motivation:

1. The importance of the project and the tests to the USMC;
2. The confidentiality of test results;
3. The importance of maintaining proper USMC decorum during all test sessions; and
4. The importance of test security.

The latter point emphasized the need to get accurate measures of performance without prior knowledge of test content. Examinees were told that the value of tests would be lessened if they told others about test content.

We also found that the use of retired Marines as Test Administrators (TAs) affected examinee motivation. As retired staff Non-Commissioned Officers (SNCOs), our TAs were well aware of proper USMC conduct; in addition, they had a rapport with the examinees that we believe insured that examinees were performing up to their capabilities.

Test Security Procedures

An important quality concern was test security. This concern was a guiding factor in planning and implementing the physical layout of the testing stations. Several measures were taken to prevent examinees from observing the performance of other Marines. For example, the indoor test stations were partitioned so that an examinee could not see any other station. Likewise, outdoor stations were well-spaced and, to the extent possible, we used terrain features to conceal each station.

Similarly, examinee movement was controlled to prevent observation and discussion. Tests were packaged so that all examinees would complete testing at the stations in approximately thirty minutes; Marines who finished early were retained at their stations until a station change was signalled. Also, during the station change, the Assistant Hands-On Manager directed traffic and prevented examinee communication.

Other test security techniques were incorporated in our TA procedures. TAs were instructed and trained to provide no information regarding test performance to examinees. This was perhaps the most difficult aspect of TA training, in that in their former USMC roles as

instructors and evaluators, their primary function had always been to train and supply feedback during testing. Both explicit and implicit feedback was held back from the examinees, both for each individual test and for the set of tests as a whole. In addition, TAs were instructed not to answer examinee questions about the quality of their performance during and after test administration. They were also trained to exercise care in scoring so that scoresheets were not visible to examinees.

A final aspect of test security was the use of alternate forms of the hands-on and Job Knowledge tests. As will be discussed in a moment, two forms of each test were developed, an "A" and a "B" version. One purpose of these alternate forms was as a test security technique: even if examinees discussed test content with to-be-tested Marines, the latter might actually receive a different version of the test.

Effects of Quality Control Procedures: Test Reliability

Effects of the quality control procedures described were assessed via both test-retest reliability and alternate forms reliability. In a strict sense, test-retest reliability is estimated by testing the same examinees twice with the same test form and then correlating the results. Parallel-forms reliability is estimated by correlating obtained scores on two parallel tests.* In the absence of proof that tests are indeed "parallel," obtained scores on alternate forms--test forms constructed with intent of being parallel--are correlated to estimate reliability of either alternate form.

The project design did not actually conform to the conditions necessary for estimating either method (test-retest or alternate forms). By having examinees take one "alternate" form (e.g., Form A) their first time through testing and the other "alternate" form (Form B) on their return, components of the two sets of conditions used to obtain the two types of estimates were essentially combined. Reliability estimates in this context, then, reflect the degree of stability or consistency of the tests, both over time and across examinees. We will focus on the relationship between examinees' obtained test and retest scores, which also happened to be scores obtained for alternate test forms.

Development of alternate forms. It was not possible to test all content in the domain of core tasks that infantrymen are expected to

* Parallel tests forms must satisfy the following assumptions (where T = true score, E = error score, X = observed score):

- (a) $X = T + E$,
- (b) expected value (population mean) of $X = T$,
- (c) true and error scores are uncorrelated,
- (d) error scores on the two test forms are uncorrelated,
- (e) error scores on one test form and true scores on the other test form are uncorrelated, and
- (f) for every population of examinees, true scores on the two forms are equal and error variances of the two forms are equal.

perform. Therefore, we devised a task selection procedure that split the allowable testing time among the various duty areas, and then randomly sampled content from the content that was testable in each of those duty areas. Maximal coverage of what we called "behavioral elements"--components of tasks that are judged similar enough in two tasks to expect that performance on the two tasks will be correlated--was the desired end product of this sampling strategy. The sampling strategy was applied once to select content from all duty areas for Form A. Content for Form B was selected by repeating the same sampling strategy with all possible content.

Some tasks were sampled for both Forms A and B. This was due in part to the fact that some duty areas had relatively few selectable tasks and in part to the random sampling process. The 12 core or basic infantry (0300) duty areas were divided into three groups on the basis of similarity of the tests within those duty areas. One group contained those duty areas with identical test content. The second group contained duty areas which had similar tests, that is, the steps were the same but the testing conditions varied to some degree. The third group contained duty areas for which the tests themselves were different. At the level of the task, of the 78 tasks sampled, 41 were common to both Forms A and B. Twenty tasks were unique to Form A, and 17 more tasks were unique to Form B.

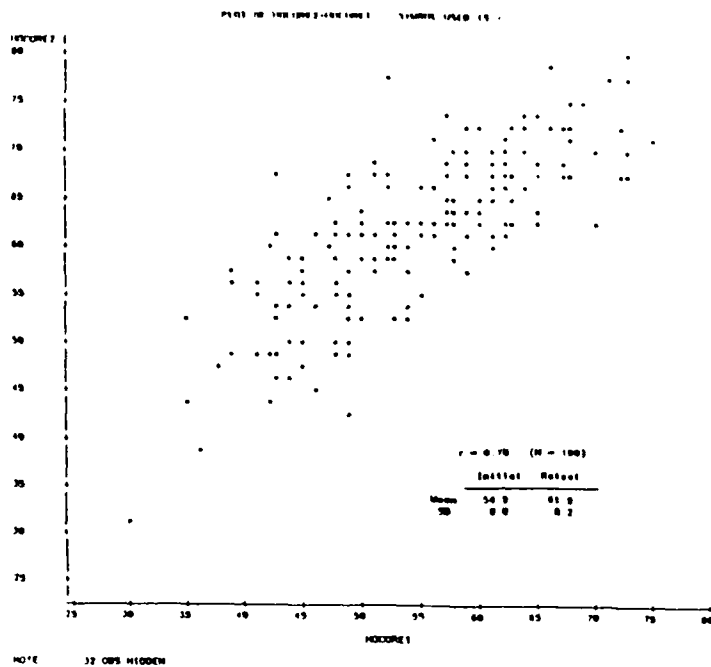
Administration. Two sets of HO measures were collected for 277 Marines at the two sites ($n=167$ at Lejeune and $n=110$ at Pendleton). However, one block of Marines at Lejeune was retested in back-to-back sessions, e.g., four days in a row. The decision was made to include in the test-retest sample those Marines who were retested within a seven to ten day interval. Thus, data for 190 0311 Rifleman Marines were included in calculating reliability estimates.

Results

Reliability estimates for the retest group and alternate form features of the project design were based on correlations obtained between Time 1 and Time 2 hands-on scores. A Pearson's r was computed to express the relationship between scores on just the core infantry content. A Pearson's r also was computed to determine the test-retest relationship for overall performance.

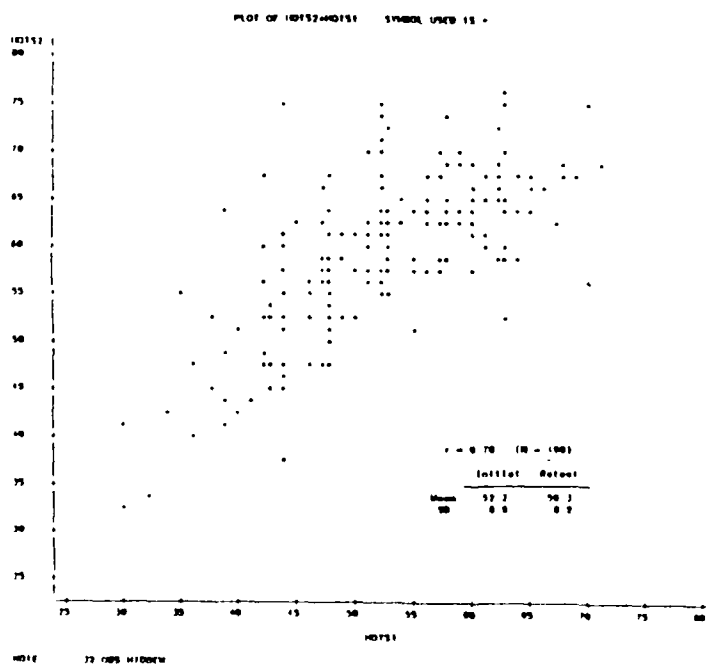
The steps taken to compute these reliabilities were as follows. A Time 1 score--called the Hands-On Core (HOCORE) score--was computed for each examinee for performance on core infantry tasks. This was a mean weighted percent correct score, obtained by applying duty area weights to mean percent duty area scores. A Time 2 HOCORE score was obtained for each examinee in the same manner. These Time 1 and Time 2 scores were plotted as shown in Figure 1. The obtained $r = .78$ ($n = 190$).

Figure 1. Alternate Form Reliability for Hands-On Test
MOS: 0311 Infantry Rifleman



Time 1 and Time 2 total competency scores were also calculated for each examinee. The total competency score, called Hands-On Total Score (HOTS) was a weighted composite of HOCORE scores and scores obtained on additional test content unique to the Rifleman (0311) MOS. The plot shown in Figure 2 illustrates the relationship: $r = .70$ ($n = 190$).

Figure 2. Alternate Form Reliability for Hands-On Test
MOS: 0311 Infantry Rifleman



Internal consistency reliability estimates also were computed for Forms A and B for the tests taken by each MOS. Table 1 gives the computed coefficient alphas and the numbers of tasks for Rifleman (0311), Machinegunner (0331), Mortarman (0341), and Assaultman (0351). Alphas ranged from .82 to .88. The number of tasks in the table differs among MOS because the number of tasks for content unique to each MOS varied. Internal consistency estimates also were computed for job knowledge tests developed to parallel the content of the hands-on tests. These alphas, given in Table 2, show the same pattern among MOS, but are slightly higher in each case. Alphas ranged from .85 to .90, with job knowledge item counts ranging from 190 to 199.

TABLE 1 Internal Consistency Reliability for Hands-on Scores				TABLE 2 Internal Consistency Reliability for Job Knowledge Test			
MOS	Form	Alpha coefficient	Number of tasks	MOS	Form	Alpha coefficient	Number of items
0311	A	0.87	71	0311	A	0.89	199
	B	0.87	68		B	0.89	199
0331	A	0.87	72	0331	A	0.88	190
	B	0.87	70		B	0.89	190
0341	A	0.86	75	0341	A	0.90	189
	B	0.88	72		B	0.89	189
0351	A	0.84	80	0351	A	0.88	190
	B	0.82	76		B	0.85	190

Conclusions

Test-retest reliability estimates were reasonable and indicate that test scores were stable over time. The internal consistency estimates support the conclusion that the tests are consistently measuring examinee performance. Our quality control procedures--design controls, addressing examinee motivation, and ensuring test security--appear to have paid off.

Validity Results for the Marine Corps Job Performance Measurement Project

**Paul W. Mayberry
Center for Naval Analyses**

The Marine Corps is in the process of conducting a long-term research effort to validate the Armed Services Vocational Aptitude Battery (ASVAB) against measures of job performance. Recently, testing for the infantry occupational field was completed. This paper examines three aspects of the Marine Corps validation research effort that may have implications for Marine Corps manpower issues:

- Validity of the ASVAB in the prediction of job performance
- Differential validity of ASVAB aptitude composites in the prediction of job performance across infantry occupational specialties
- Interaction of aptitude and experience in the prediction of job performance.

ASVAB Validity

The Marine Corps initially selected the infantry occupational field because it is the largest and is often the most controversial in terms of establishing prerequisite aptitude requirements. In total, over 2500 Marines were tested in five military occupational specialties (MOSs). The findings for each MOS were particularly noteworthy – ASVAB was found to be highly valid as a predictor of hands-on performance. Validity coefficients for the prediction of hands-on performance by the General Technical (GT) aptitude composite ranged from 0.48 to 0.68, see table 1. Such results contradict the common belief that anyone can function as an infantryman; in fact, aptitude is very relevant to the successful performance of infantry job requirements. These findings strongly support the continued use of ASVAB in the selection and classification of recruits.

A second significant outcome relates to the second-term unit leaders. The magnitude of the validity coefficient for this group also indicates the robust ability of aptitude to predict second-term performance. Again, those individuals in the second-term with higher aptitudes tend to become better performers. This outcome has significant implications for the types of individuals that the Marine Corps should recruit so as to be able to maintain qualified personnel in future years to man its leadership positions.

TABLE 1
VALIDITY OF ASVAB IN PREDICTING
HANDS-ON PERFORMANCE

MOS	Validity			
	GT		AFQT	
	Sample	Corrected	Sample	Corrected
0311 Rifleman	0.47	0.62	0.40	0.55
0331 Machinegunner	0.50	0.68	0.49	0.66
0341 Mortarman	0.42	0.48	0.33	0.38
0351 Assaultman	0.41	0.50	0.38	0.46
0369 Second-term unit leader	0.48	0.60	0.41	0.50

Differential Validity of ASVAB Aptitude Composites

The ASVAB is composed of 10 subtests which the Marine Corps combines into 4 aptitude composites that are used to classify recruits into occupational specialties. The extent to which these composites can differentially predict performance across all Marine Corps jobs is paramount to the use of ASVAB and the proper allocation of personnel into areas for which they have the greatest potential to perform.

The four Marine Corps aptitude composites are: general technical (GT), mechanical maintenance (MM), electronics repair (EL), and clerical/administrative (CL). (Although the AFQT is also an aptitude composite, it is not explicitly used by the Marine Corps for selection and classification purposes; rather, it is a quality indicator used for tracking trends in accessions.) The GT composite is used for classification of recruits into infantry specialties.

Figure 1 plots the corrected validities for each infantry specialty by each of the aptitude composites. Despite the relatively high magnitude of these validities, the MM composite is a better predictor of infantry performance, particularly for the rifleman (0311) and mortarman (0341) specialties. The EL composite is comparable to MM for the machinegunners (0331) and assaultmen (0351). In all cases, the CL composite is substantially lower, implying that it is only a mediocre predictor of infantry performance. These results provide empirical support for differential validity

by aptitude composite and even slight differentiation across specialties. Also, the findings indicate that the Marine Corps could significantly benefit by using MM to classify recruits into the infantry specialties. However, before implementing such findings, additional investigation is required to determine if any possible unacceptable implications would result.

In addition to the hands-on performance tests, the Marine Corps also administered a written job knowledge test (JKT). This test was parallel in content to the hands-on tests but administered in paper-and-pencil form. Using the JKT as the performance criterion resulted in higher overall validities, primarily due to the common testing medium (i.e., both the ASVAB and JKT are written tests). The validities of the aptitude composites against the JKT scores are shown in figure 2. It is interesting to note the lack of differential validity by the four aptitude composites using JKT as the performance criterion. Contrasted to the findings for the hands-on test, the MM composite is now the worse composite for three specialties and CL predicts infantry performance almost as well as the other composites. Given that the ASVAB has historically been validated against training grades and that training grades are based on tests similar to the JKT, it is not surprising that researchers have found the ASVAB to have limited differential predictive validity. Hands-on performance has been defined by the Joint Service Job Performance Measurement (JPM) Working Group as the benchmark performance criterion against which the ASVAB should be validated. These findings demonstrate that indeed the ASVAB does have differential validity against a high fidelity performance measure.

Interaction of Aptitude and Hands-on Performance

The relationship between hands-on performance tests and job experience is important for establishing the measurement validity of hands-on tests. The expectation is that hands-on test performance should increase with experience as job incumbents acquire higher levels of proficiency through on-the-job training and more advanced instruction. Also, to the extent that aptitude is a valid predictor of performance, personnel with high aptitude should out-perform their counterparts with low aptitude. The interesting question involves the interaction of both of these variables and the prediction of performance: do high and low aptitude individuals differ in their performance across all levels of experience or can experience compensate for lower aptitude? The Marine Corps infantry project examined this aptitude-experience interaction by investigating three "indicators" of experience: time in service, pay grade, and recency of task performance.

Figure 3 plots the mean hands-on performance for both high and low AFQT personnel (I-III A versus IIB-IV) at various time-in-service intervals for the four first-term infantry specialties. The results indicate that performance differences

between high and low AFQT individuals were generally large for inexperienced personnel, but that these performance differences tend to lessen as time in service increases. The trend towards decreasing performance differences over time in service has been found in the Marine Corps initial JPM study and by some of the other Services as well. The finding has typically been explained by either a ceiling effect on the test or that higher aptitude personnel are promoted to supervisor positions and no longer perform the tasks on which they were tested. Neither of these explanations appear reasonable given the Marine Corps investigation of the other two indicators of experience. In addition, the Marine Corps specifically designed its tests so that examinees were tested on job requirements beyond their level of responsibility so as to preclude the possibility of a ceiling effect.

When examining the performance differences between high and low aptitude personnel using pay grade as the experience measure, such leveling-off trends are nonexistent. Figure 4 illustrates this continuously improving performance trend. The performance of Marines in pay grades E4-E5 is superior to Marines in pay grades E3 and E1-E2, and high aptitude personnel are always better performers than low aptitude personnel. Thus, it is evident that the Marine Corps promotion system is properly advancing its higher performers. It appears that the time-in-service definition of experience is contaminated at high levels by low performing personnel who have not been promoted to the ranks that would be consistent with their time in service.

The final indicator of experience - recency of task performance - has possible implications for addressing the interaction of aptitude and frequency of training. Is training able to compensate for lower aptitude in the prediction of job performance? Loosely translating training as the recency of task performance, figure 5 shows the tradeoffs between aptitude and performance recency. Given comparable levels of recency of task performance, high aptitude personnel always significantly outperform low aptitude personnel. In fact, low aptitude persons with recent task performance (less than six months) are comparable to high aptitude persons who have only limited task performance (greater than six months). The same is true for low aptitude persons with limited task performance being comparable to high aptitude persons with only task instruction but no opportunity to perform the task. Such findings speak strongly to the need for high quality personnel because training or refresher task performance is not always possible in time of conflict.

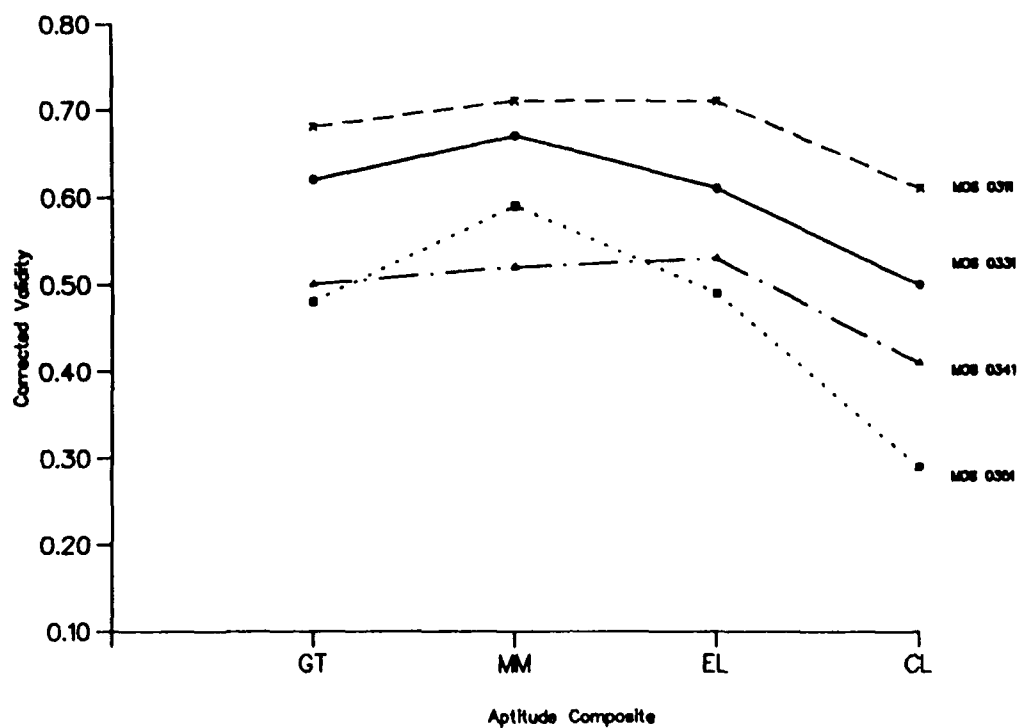


FIG. 1: CORRECTED VALIDITIES OF HANDS-ON TOTAL SCORE AND APTITUDE COMPOSITE SCORE BY MOS

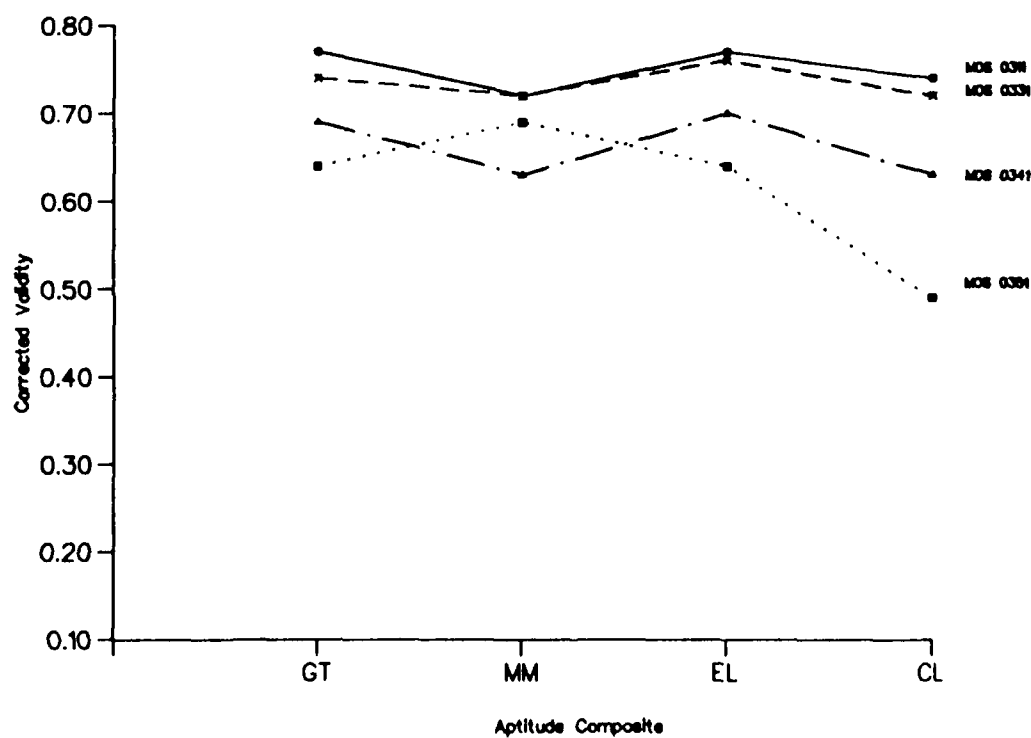


FIG. 2: CORRECTED VALIDITIES OF JOB KNOWLEDGE TEST AND APTITUDE COMPOSITE SCORE BY MOS

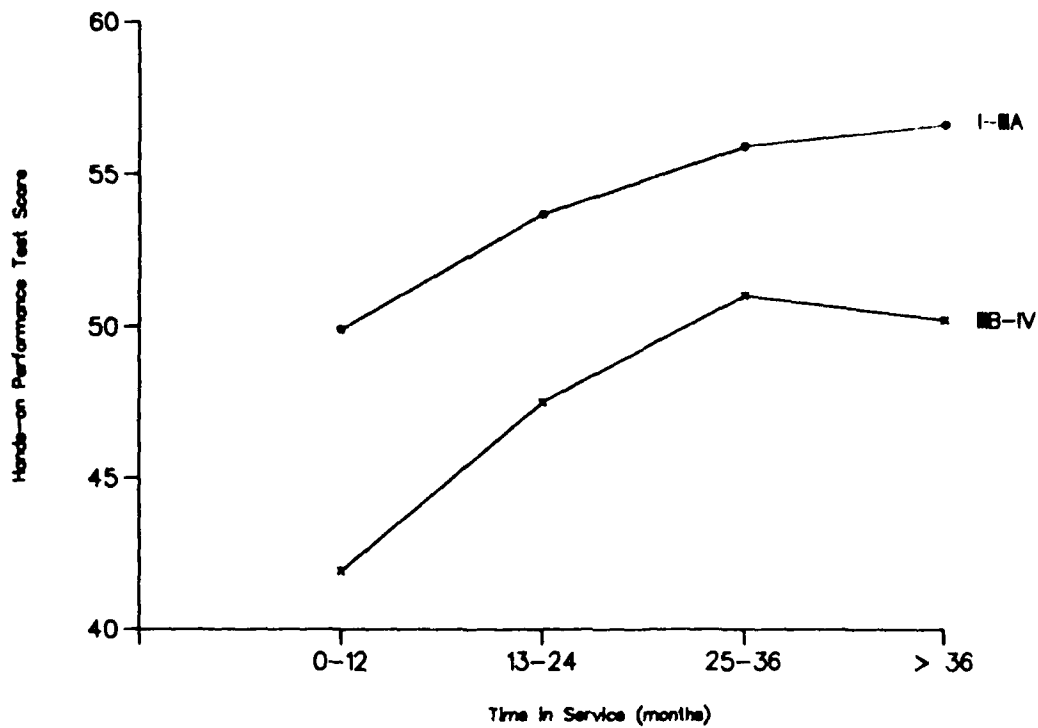


FIG. 3: MEAN HOPT SCORES BY AFQT AND TIME IN SERVICE FOR INFANTRY RIFLEMAN

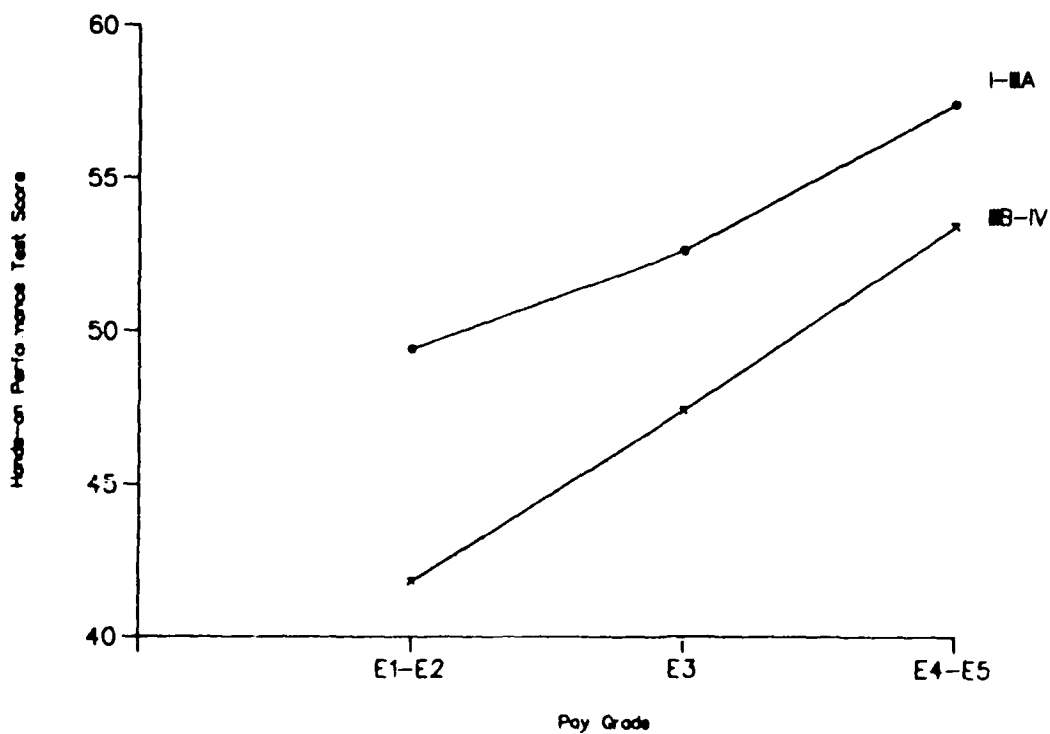


FIG. 4: MEAN HOPT SCORES BY AFQT AND PAY GRADE FOR INFANTRY RIFLEMAN

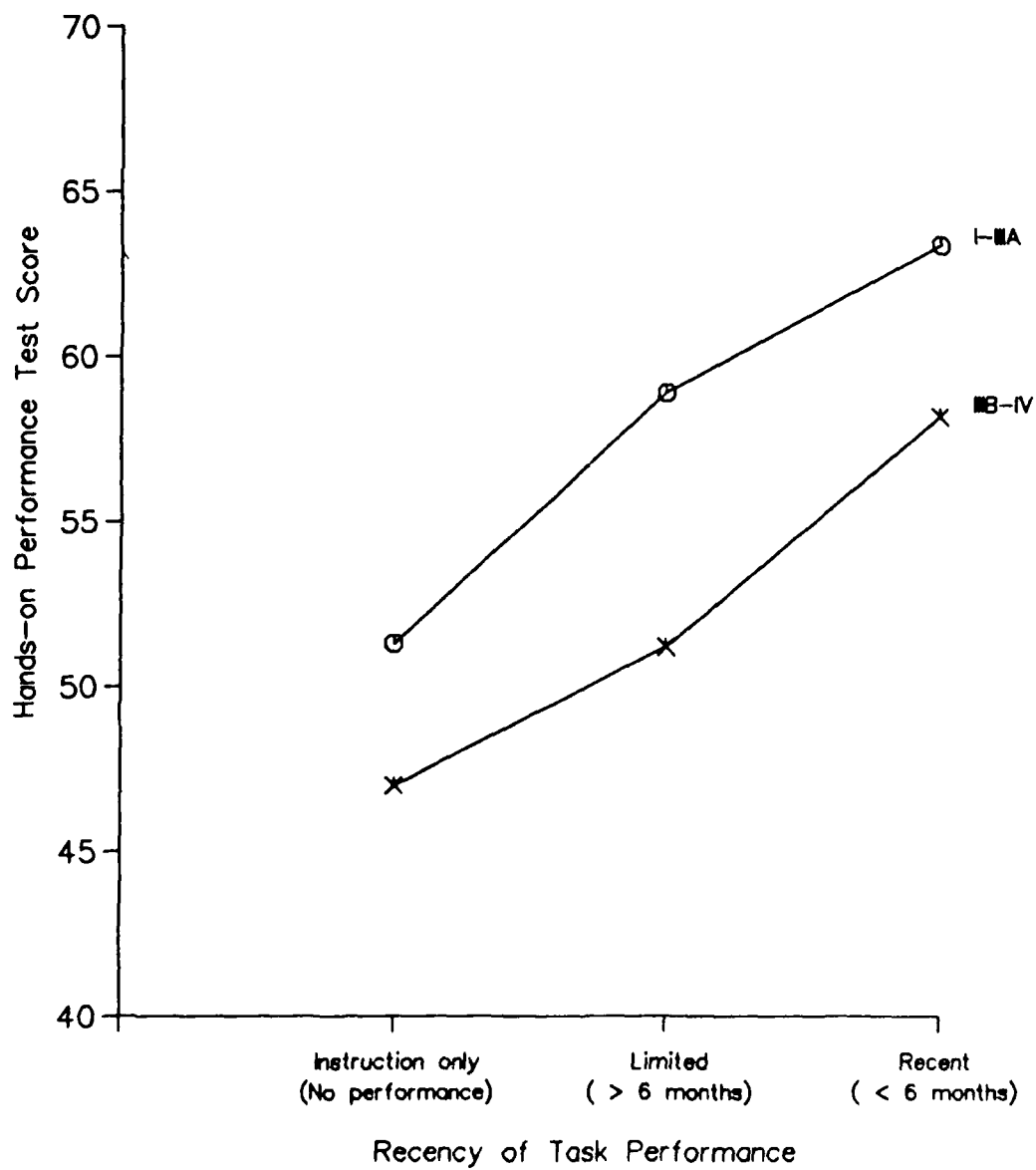


FIG. 5: MEAN HOPT SCORES BY AFQT AND RECENCY OF TASK PERFORMANCE

EVALUATING PROJECT A TESTS FOR SELECTING GUNNERS

Clinton B. Walker
U.S. Army Research Institute
Alexandria, Virginia

In December 1987, the U. S. Army Training and Doctrine Command started two programs for trainees in gunnery specialties: Skills Selection and Sustainment (S3) for selecting and training soldiers who have aptitudes for gunnery, and Excellence Tracking for advancing promising trainees rapidly toward positions of leadership. For selection into these programs, Army Research Institute provided new tests from Project A: computerized tracking tests; pencil-and-paper tests of spatial abilities; and a temperament battery. The three papers in this panel describe the initial results of this testing, including the incremental validity of the new tests relative to ASVAB.

Scott Graham evaluates the utility of the test battery for selection into an accelerated track in Armor training (Excellence in Armor). High validities emerge for the spatial/psychomotor tests against the simulator of tank gunnery in a sample of over 500 trainees, differences between Excellence and normal track soldiers in gunnery are described, and the relation of ABLE scores to dropout from the Excellence program is reported.

Elizabeth Smith and Martin Walker describe the use of a spatial/psychomotor composite to predict performance on the M-70 TOW missile simulator. In 326 trainees, those meeting a given cut score qualify faster and attain higher qualifying scores than those not meeting it. Scores on ABLE closely resemble data from Project A.

Ilene Gast and David Johnson report correlations of the spatial and psychomotor tests with performance on the Realistic Air Defense Engagement Simulator (RADES). The results vary, probably due to unique conditions of performance in air defense.

SELECTING SOLDIERS FOR THE EXCELLENCE IN ARMOR PROGRAM

Scott E. Graham
U. S. Army Research Institute
Fort Knox Field Unit

Success on the modern battlefield requires soldiers who can maximize the capabilities of their weapon systems and who are excellent leaders. To help meet these needs, the U.S. Army Armor Center (USAARMC) in 1984 initiated the Excellence in Armor (EIA) program as a complement of initial-entry training. The EIA program identifies early-on in One Station Unit Training (OSUT) high-quality, motivated soldiers. The selected soldiers receive accelerated training on hard-skill armor tasks, with successful performance resulting in early promotions. The goals of the EIA program include increased retention of high-quality enlisted personnel and accelerated progression of EIA graduates into tank commander (TC) assignments.

Reports from field commanders indicate the EIA program is presently successful, in that EIA graduates are superior to other OSUT graduates. As part of continuing efforts to improve the Armor force, USAARMC is interested in enhancing the quality of its EIA graduates through improved EIA selection tests. The research reported here examines a potential new set of EIA selection tests and criteria. The work was done as technical advisory service to USAARMC and was conducted under the auspices of the U. S. Army Training and Doctrine Command's Skills Selection and Sustainment (S³) program.

The S³ selection tests were developed by the U. S. Army Research Institute and were designed to measure spatial, psychomotor, and leadership abilities. Should the implementation of the S³ tests prove effective, the result would be an EIA graduate population with even greater leadership and warfighting abilities.

Purpose of Research

The purpose of the research was to assess the utility and impact of the S³ predictors tests as additional selection criteria for the EIA program. The major questions addressed by the research were:

1. Do the S³ tests predict simulated tank gunnery performance?
2. Do soldiers currently selected for EIA differ from those not selected on the S³ predictor tests?
3. How would S³-based selections differ from current selections?

TEST MATERIALS

The S³ selection battery included two paper-and pencil tests, the orientation test and maze test, and two computerized tracking tests. Scores from these tests were combined into a spatial/-psychomotor composite. The predictor battery also included the Assessment of Basic Life Experiences (ABLE) Short Form, a temperament scale designed to assess leadership potential. The predictor tests are described elsewhere in this volume.

A tank gunnery criterion test was developed and administered on the M1 Institutional-Conduct of Fire Trainer (I-COFT). The I-COFT is a high-fidelity tank gunnery simulator which has become a central component in the suite of armor gunnery training devices. TC and gunner controls on the M1 I-COFT are virtually identical to those in the actual tank, making the I-COFT analogous to flight simulators used in military and commercial training. Recently the I-COFT has begun to be used as a device for measuring tank gunnery proficiency.

The I-COFT gunner's test developed for this research contained four exercises taken from the I-COFT's Target Engagement Practice Exercises (TEPE). The selected exercises included offensive and defensive engagements fired with daylight and thermal sights under normal and degraded operational conditions.

Two performance measures were obtained from each exercise: percent hits and opening time. Percent hits was simply the number of targets hit divided by the number of targets presented. Opening time measured the amount of time from when a target appeared until the first round was fired. A speed/accuracy composite score was also computed by subtracting the standardized opening time from the standardized percent hits. The speed-/accuracy composite was then transformed into a t-score.

METHOD

Four hundred seventy-nine 19K (M1 tank) OSUT soldiers completed both the predictor battery and the I-COFT gunnery criterion test. The soldiers were from five training companies of the 1st Armored Training Brigade (1ATB), Fort Knox KY. Predictor Scores were also obtained from an additional 1143 OSUT soldiers, including MOS 19D (scouts) and 19E (M60 tank), for a total of 1642.

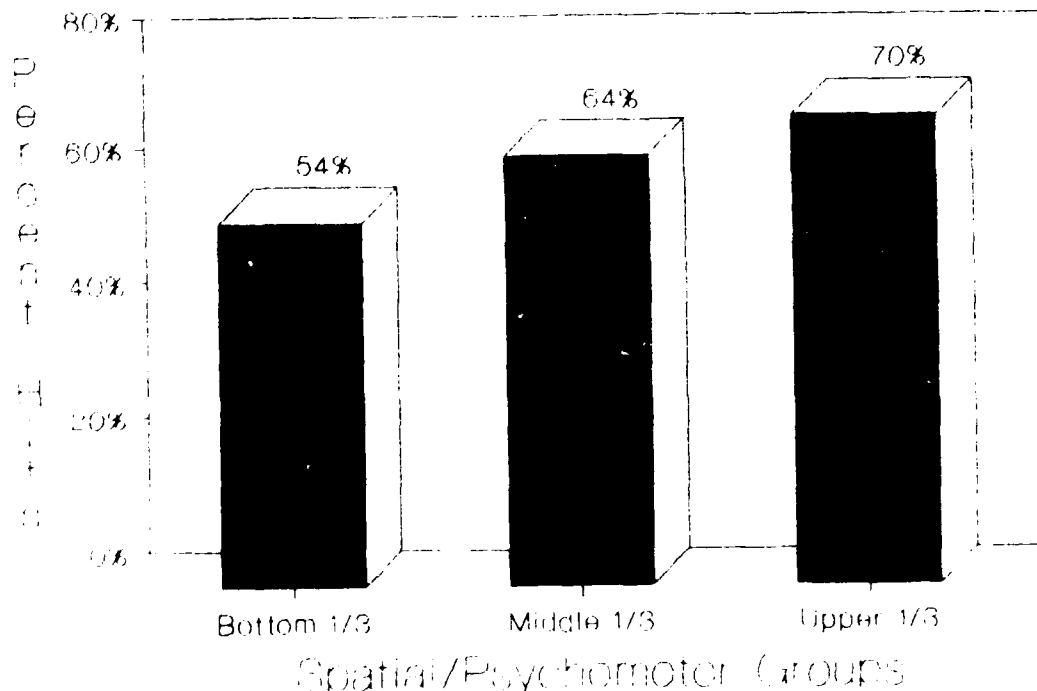
The soldiers took the predictor tests at the Fort Knox Reception Station on the third day of their initial processing into the Army. The I-COFT tests were administered by I-COFT Instructor/Operators (I/O) to the 19K soldiers during the last (or 20th) hour of OSUT I-COFT training; this fell in the tenth week of OSUT training. The 19K soldiers selected for the EIA program received an additional 14 hours I-COFT training and were readministered the I-COFT test at the end of this training.

RESULTS

S³ Tests and I-COFT

Performance on the S³ predictor battery was highly correlated with I-COFT gunnery performance. Specifically, the correlation of the spatial/psychomotor composite with I-COFT percent hits was .48; with opening times, -.52; and with the speed/accuracy composite, .54. The soldiers were then split into three equal sized groups as a function of their spatial/psychomotor composite score.

The figure below shows I-COFT percent hits for the lower, middle, and upper spatial/psychomotor groups. An Analysis of Variance (ANOVA) found the differences in means to be significant with $F(2,476) = 51.9$, $p < .0001$. The same pattern was found for opening times and the speed/accuracy composite. For all three I-COFT performance measures, the difference between the upper and lower spatial/psychomotor groups was greater than one standard deviation.



Regression Analyses

A series of stepwise regression analyses were conducted to predict I-COFT performance from the S³ scores and other available predictor information. For each of the regression analyses, the criterion measure was COFT speed/accuracy. The table below shows the zero-order correlations of the predictors with the I-COFT composite.

Predictor	(n = 446)	r
Spatial/psychomotor composite		.54
I-COFT training score		.48
Tracking test score		.46
Spatial test score		.40
GT score		.34
ABLE		.11

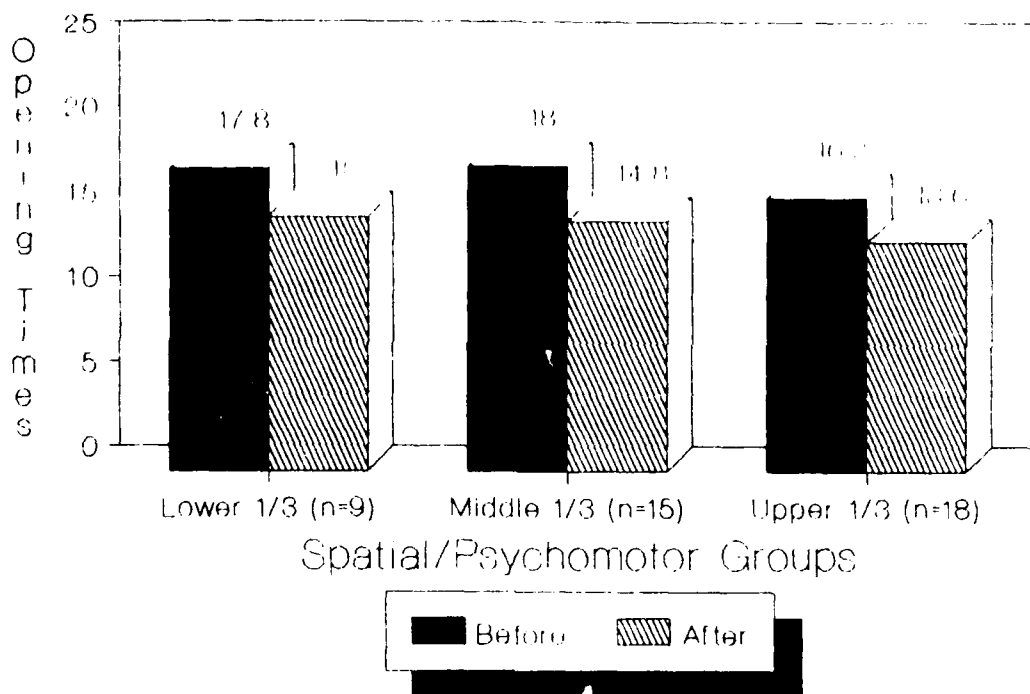
The results of the stepwise regression analyses found the spatial/psychomotor composite and the I-COFT training score combined to be the best predictor of I-COFT speed/accuracy, yielding a multiple R of .59. The I-COFT training score, developed for this research, measured I-COFT training performance during OSUT. A separate regression analysis showed that the I-COFT training score and GT did nearly as well as the S³ tests in predicting I-COFT speed/accuracy, producing a multiple R of .52.

Analyses Following Additional EIA I-COFT Training

Soldiers selected for EIA received an additional 14 hours of I-COFT training as part of the accelerated EIA training. Not surprisingly, the additional hours of I-COFT training lead to a marked improvement in gunnery performance, both in terms of speed and accuracy, e.g. percent hits increased from 71% to 84%.

The figure below shows the mean opening times for the EIA soldiers separated into the lower, middle, and upper spatial/psychomotor composite groups. The cut scores for the three groups are the same as those used in the earlier analyses on the entire 19K OSUT sample. Separate ANOVAs found significant differences both before the additional hours of training, $F(2,37) = 3.53$, $p < .05$, and after $F(2,39) = 3.23$, $p < .05$. For the before training ANOVA, the n's were 9, 14, and 17.

The data show the S³ tests to be equally predictive of I-COFT performance at the two different levels of training. A similar finding was that the tracking test correlated with I-COFT speed/accuracy .37 before the additional training and .35 after the additional training. These correlations are lower than those found for the entire OSUT sample because of restricted range, i.e., the EIA soldiers were generally from the top of the I-COFT performance distribution.



S³ Test Scores of EIA and Normal Track Soldiers

Analyses comparing the S³ test scores of soldiers currently selected for EIA to those not selected found the EIA soldiers scored significantly higher on all of the tests. In addition, the EIA soldiers had higher GT scores, an ASVAB measure related to mental ability. A discriminant analysis further supported the point that the EIA selection process was already, to some extent, selecting soldiers on the S³ and GT dimensions. Each of the tests correlated highly with the single discriminant function; the canonical correlation was, however, a rather low .24.

S³ Overlap with Current EIA Selection

The current EIA selection process selects approximately 11% of OSUT soldiers for the program. As just shown, soldiers currently selected for the EIA program score higher on the S³ tests than those soldiers not selected. It may be that the 11% of the soldiers who are selected for EIA are also the top scorers on the S³ tests. A comparison of the two procedures, however, found only a moderate overlap. For example, if selection into EIA required the soldier to be in the upper 30% of the S³ distribution, 52% of the currently selected EIA soldiers would be above the cut.

As indicated by the low canonical correlation in the discriminant analysis, the current EIA selection procedure is not maximizing the selection of soldiers on the basis of psychomotor leadership, and mental abilities. Ongoing research is monitoring the EIA selection process with the goal being to determine those dimensions that are being used to select the EIA soldiers.

DISCUSSION

The results clearly demonstrate that the S^3 spatial and psychomotor tests are valid predictors of tank gunnery performance, as measured on the I-COFT. Not only were the correlations high, but the 2 1/2 month interval between the predictor and criterion tests suggests the relationship should remain stable over time. Furthermore, the strength of the relationship shrunk only slightly when the EIA soldiers were given considerable additional training. Taken together, the data show the S^3 tests to be valid over time and varying levels of gunnery proficiency.

The comparison of soldiers selected for EIA to those not selected found the EIA soldiers had better spatial, psychomotor, and leadership skills. The EIA soldiers also had higher GT scores and performed better on the I-COFT gunnery test. The analyses indicated, however, that including the S^3 test scores in the EIA selection process would result in EIA soldiers with stronger gunnery skills. The analyses also showed that I-COFT training data were not effectively used in the selection process. Modifications to the OSUT training schedule might have to be made to do so.

The current EIA selection procedure relies heavily on supervisory evaluations which are believed to be good measures of soldier motivation and leader potential. Subjective appraisals are a necessary part of the EIA selection process, but the analyses suggest there is room for improvement. Less than five percent of the variance discriminating EIA soldier selection was accounted for by the S^3 tests, GT scores, and I-COFT training performance data. Given that the tests measure psychomotor, spatial, leadership, and mental abilities, plus hands-on I-COFT training performance, one would expect a greater difference between EIA and normal track soldiers in these important areas.

The critical question concerning the utility of the S^3 tests as an EIA selection tool remains unanswered. Would the overall quality of EIA graduates be higher if certain soldiers selected for the accelerated training program, i.e., those who scored low on the S^3 tests, were replaced by soldiers who scored high on the test? The data suggest that the quality of the EIA graduates would increase if the S^3 were used in the selection process. The result would be EIA soldiers with stronger leadership and warfighting abilities.

Testing Psychomotor and Spatial Abilities to Improve Selection of TOW Gunners¹

Elizabeth P. Smith²
U.S. Army Research Institute
Alexandria, Virginia

Martin R. Walker
U.S. Army TRADOC Analysis Command
FT Benjamin Harrison, Indiana

The emergence of highly positive results from the Army's Project A to improve enlisted selection and classification, -- especially in relation to gunner performance with M-1 tanks (Smith & Graham, 1987) and Tube-launched Optically-tracked Wire-guided (TOW) missiles (Grafton, Czarnolewski, & Smith, 1988) -- heightened interest in quickly implementing the new paper-and-pencil and computerized tests. A decision was made to use two of the ten computer tests (One Hand and Two Hand Tracking) and two of the six spatial tests to form a psychomotor/spatial (PScomp) composite to be used in selecting gunners for a number of Military Occupational Specialties (MOS). Reexamination of data from Smith & Graham (1987) and Grafton, et al. (1988) indicated this composite validly predicted performance on two simulators, the M-1 Unit Conduct of Fire Trainer (UCOFT) and M-70 TOW simulator. Also, a shortened version of the Assessment of Background and Life Experiences (ABLE) was proposed as a screen for assignment to "Fast Track" or accelerated training programs.

This paper examines results from the baseline period of implementation (prior to actual use of the scores) of these psychomotor and spatial and ABLE for assignment of TOW gunners. In addition, the paper looks at some results from a live-fire TOW test conducted by the US Army Infantry Board (USAIB), Fort Benning, Georgia. The purpose of the live-fire TOW test was to evaluate a new TOW training strategy that incorporated the use of a Multiple Integrated Laser Engagement Simulator (MILES) training device with the current M70 Program Of Instruction (POI) (USAIB, 1988). Thus, the purpose of this paper is to examine (a) the relationship of psychomotor/spatial ability to performance for TOW gunners in simulated and live fire, (b) the effects of different cut scores on the distribution of soldiers by gunner performance qualifications, and (c) performance on the ABLE.

Conditions for this research differed from Grafton, et al. (1988) in several ways. Predictor tests were administered by military personnel in an operational rather than research setting. Only four tests were given and any "warm up" effect of previously presented computerized tests was eliminated. Changes were made in procedures for training and qualifying on the M70 by the Infantry Center as well. Soldiers could qualify after fewer training trials, had to pass on only one rather than two trials, and, for the majority of the sample, qualified with a TOW mounted on a High Mobility Multipurpose Wheeled

¹The views expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Institute or the Department of the Army.

²Now employed at the Justice Department, Washington, DC.

Vehicle (HMMWV) rather than ground mounted. Despite these differences, the findings here were expected to replicate those of the earlier research.

METHOD

Sample

Predictor data were collected for over 1512 Infantry trainees from February - June 1988 in the generic 11X MOS. Of this sample, 326 trainees subsequently were assigned to MOS 11H (TOW Gunner) training and obtained M70 simulator scores. A subsample of 60 soldiers was randomly selected to participate in a live-fire TOW test conducted by the USAIB from 6-22 June 1988.

Variables

Predictors. Two computer tracking tests measuring psychomotor precision (One Hand Tracking) and multilimb coordination (Two Hand Tracking) provided a single standardized score. This score was weighted double and added to a single standardized total score over tests measuring spatial orientation (Orientation) and scanning (Maze) to form PScomp. The General Technical Score (GT) from the Armed Services Vocational Aptitude Battery (ASVAB) was used as a second predictor when it was available from records. The ABLE Short Form score, a measure of discipline and leadership, is the total over 68 biodata items scored from 1 to 3 points each. Additional items provide scores for Non-random Responding (i.e. failing to answer seriously) and Unlikely Virtues (i.e., "faking good"). Both PScomp and ABLE scores were standardized using the Project A Longitudinal Validation sample from Combat MOS as the norm group.

Criteria. The M70 score is a measure of tracking ability summed over ten simulated TOW firings. The score for each firing signifies the proportion of time-on-target during tracking and indicates likelihood of a target hit. The maximum total score is 1000 (i.e., 10 times 100%). Gunner qualification categories are assigned from M70 scores as follows: "Unqualified" (score <550), "2nd Class" (550-649), "1st Class" (650-749) and "Expert" (750-1000). The primary simulation criterion was the qualifying score of record (M70 Score), which varied by number of trials to attain. Two secondary criteria were examined: the score attained on Trial 3 and the number of trials required to qualify. For the live-fire test, the primary criterion was the probability of first round hit (Ph) with the TOW missile.

Procedures

Soldiers were tested on the predictor tests in small groups in one 2-hr block in their first week in One Station Unit Training (OSUT). Criterion data were obtained during the normal training program about two months later. After two training trials, soldiers were able to qualify by obtaining a score of 550 or greater on Trial 3 or on up to the seventh trial. For the live-fire test, each soldier fired one live TOW missile after completing a prescribed training POI. Because of USAIB's conclusion of no differences due to POI, the data were not analyzed separately for different POI.

Analyses

Means and standard deviations were calculated for both the 11H and 11X samples on the predictors. Equivalency tests were used to compare predictor scores of those selected and not selected under the current practices. Correlations between predictors and criteria were computed, and criteria were regressed on PScomp plus GT. Next, to examine Gunner Qualifications as a function of different selection rules, we selected two sample cut-off scores, 61 and 68. This enabled comparison of outcomes based on the current practices and those based on Project A tests. Data from the live-fire test grouped using a cut-off score of 60 on the predictor test were compared using Chi square.

RESULTS

Table 1 presents means and standard deviations for the TOW and non-TOW trainee samples. T-tests indicated only one difference (total Tracking) was statistically significant ($t_{1442} = 2.62$, $p < .009$) in favor of the 11H sample, but this is not likely a meaningful difference. Given that Tracking and Spatial scores are standard (T) scores, the means here indicate above-average Tracking ability in comparison to the Project A norm group, yet this may be a function of greater effort rather than greater ability, due to the operational conditions. Of particular interest are the ABLE Unlikely Virtues and Non-Random Response scores which are nearly identical not only with each other, but also with data from Project A (Project A, 1986), indicating no effect due to experimental vs. operational conditions.

Table 1

Means and Standard Deviations for TOW and Non-TOW Infantry Trainees

Variable	11X			11H		
	<u>n</u>	<u>M</u>	<u>sd</u>	<u>n</u>	<u>M</u>	<u>sd</u>
Tracking ^a	1186	53.46	8.34	257	54.96	8.03
Spatial	1186	50.07	9.01	259	50.26	8.18
PScomp	1163	157.18	22.36	259	160.37	19.92
Percentile ^b	1163	60.00	-	326	60.87	24.88
ABLE	1186	164.70	16.48	259	166.92	14.82
Unlikely Virtue	1186	16.13	2.57	259	16.03	2.30
Non-random	1186	7.7	5.57	259	7.76	.62
General Technical	-	-	-	326	106.07	11.78
Trial 3				326	626.52	130.12
M70 Score				326	661.25	79.24
No. Tables				326	3.27	.73

^at significant at $p < .01$

^bPercentile estimated for 11H.

Correlations for PScomp and GT with the three simulation criteria are given in Table 2. Correlations are greatest for Trial 3, which is likely to be the best criterion from the standpoint that each soldier has had the same amount of training at that time. Number of trials is the least effective criterion in that approximately 84% of the sample of 326 qualified on Trial 3. Table 2 also indicates that little increase in validity is achieved by regressing the criteria on both PScomp and GT, which correlate .31 with one another. Indeed, GT does not contribute significantly to the regression models for either M70 Score or Trial 3.

Table 2

Correlations Between Ability and Criterion Scores for 11H

Variable	Trial 3	M70 Score	No. of Trials
PScomp	.37	.27	-.23
GT	.29	.15	-.22
PScomp, GT ^a	.38	.28	-.28

Note: $p = .0001$ except for $r_{gt, m70} p = .007$

^aMultiple R from regression.

Table 3 summarizes data from the live-fire test. Although there was a demonstrated difference in first round Ph of 12% for those scoring above and below the cut-off, based on this small sample, the difference was not statistically significant ($\chi^2_{59} = 1.24$, $p = 0.26$).

Table 3

TOW Live-fire Test Results by Predictor Score Category

PScomp Score	Attempts	Misses	Probability of hit(Ph)
At or above 60	33	5	.85
Below 60	26	7	.73
All soldiers	59	12	.80

Figure 1a compares distributions across Gunner Qualifications for the total 11H sample (selected using current procedures) and for samples derived using cut scores of 68 and 61 on PScomp. The percent of change within the three qualifying categories is depicted in Figure 1b. Clearly, applying

PScomp for selection results in higher percentages of Experts and 1st Class while reducing the number of 2nd Class and unqualified soldiers. The predictor test at the two cut-off scores would have screened out approximately 54% and 44% of the trainees that were selected for MOS 11H. However, 42% of the 1512 soldiers tested scored above 68 and more than half scored above 61 on the predictor test. Therefore, even at the 68 cut-off, more than 645 soldiers scored high enough to fill the 326 MOS 11H training seats. In addition, comparisons of mean M70 Scores for groups above and below the cut-offs were significant ($t_{324} = 4.38$ and 4.72 , $p = .0001$). The 38-40 point difference was such that the mean fell into the 1st Class for the selected and 2nd Class category for the non-selected. Similarly, the mean Number of Trials required to qualify by the two groups differed significantly ($t_{281} = -4.57$, $p = .0001$ and $t_{249} = -3.46$, $p = .0006$). Finally, Figures 2a and 2b demonstrate the reduction in extra trials (and time) to qualify by implementing the PScomp testing. At the 68 cut-off, 95% qualify at Trial 3 and at the 61 cutoff, 91% do, as compared to 84% of the original sample.

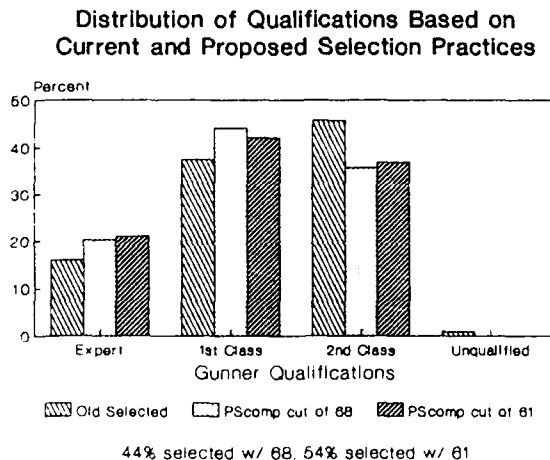


Figure 1a

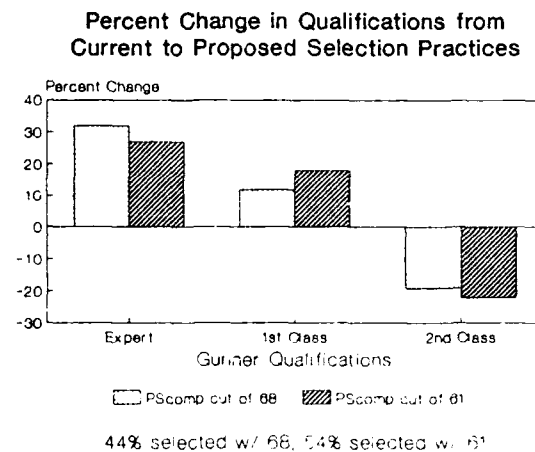


Figure 1b

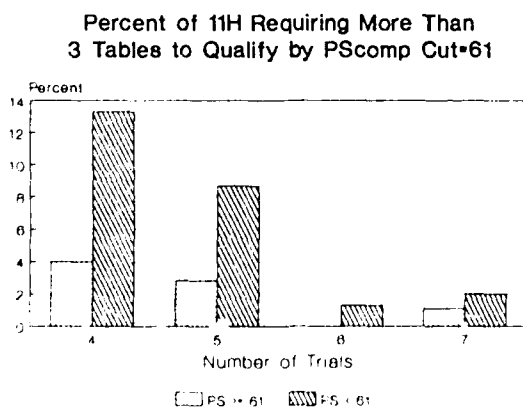


Figure 2a

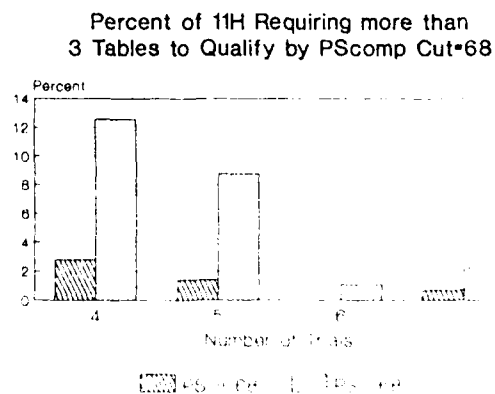


Figure 2b

A final consideration was how excluding trainees on the basis of a low ABLE score (less than 10th percentile) would affect gunner selection. In the present data, doing so resulted in disqualifying only an additional 3 of the 259 soldiers with ABLE scores.

CONCLUSIONS

Of critical importance to the implementation of new procedures is evidence that they will yield improvements. An important aspect of this evaluation was that it crossvalidated the predictor test in both a training and a pseudo-combat environment. Re-analysis of the earlier TOW data (Grafton, et al., 1988) using the Pscomp composite yielded a correlation of .31 with M70 Score. The .28 obtained here is only slightly less. If we consider the correlation for Trial 3, which is perhaps a better criterion, the relationship is stronger ($r = .37$). The data have consistently shown that by using these tests of psychomotor and spatial ability, it is possible to select gunners who not only are more proficient, but who also take less time to train and qualify. Clearly, applying PScomp for selection results in higher percentages of Experts and 1st Class while reducing the number of 2nd Class and unqualified soldiers. Similarly, the majority of the soldiers who needed extra trials, and, hence, extra time and trainer effort, would have been disqualified from becoming 11H gunners. Also, performance on the ABLE tends to parallel that of the research samples. This indicates that we can expect ABLE to predict "Will Do" kinds of performance important for soldiers selected for specialized "Fast Track" courses of training.

Countering these arguments in favor of the testing program are those that contend with the increased cost and effort required to conduct the testing. Start-up costs are greatest and should diminish over time. The actual cost-effectiveness of implementation of the testing program remains to be evaluated.

REFERENCES

- Grafton, F. C., Czarnolewski, M. Y., & Smith, E. P. (1988). Relationship between Project A psychomotor and spatial tests and TOW gunnery performance: A preliminary investigation. SCTA Working Paper RS-WP-87-10. Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Project A (March, 1986). Concurrent Validation data presented to the Project A Scientific Advisory Group.
- Smith, E. P. & Graham, S. E. (1987). Validation of psychomotor and perceptual predictors of Armor Officer M-1 gunnery performance. (ARI Technical Report No. 766). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- U.S. Army Infantry Board (USAIB) (1988). Concept evaluation test program (CEP Test) of Tube-launched Optically-tracked Wire-guided (TOW) Training Strategy. USAIB Project No. 3901, U.S. Army Infantry Board, Fort Benning, GA.

Evaluating Psychomotor and Spatial Tests for Selecting Air Defense Gunners¹

Ilene F. Gast

U.S. Army Research Institute for the Behavioral and Social Sciences²

David M. Johnson

U.S. Army Research Institute for the Behavioral
and Social Sciences, Fort Bliss

The U.S. Army Research Institute evaluated the use of spatial and psychomotor tests, developed for Project A: Improving the Selection and Classification of Enlisted Personnel (Eaton, Goer, Harris & Zook, 1984), to predict detection, identification, and engagement of aerial targets in two Air Defense occupations. Included were 16P personnel who operate the Chaparral weapon system, a vehicle-mounted heat-seeking missile, and 16S personnel who operate the Stinger weapon system, a shoulder mounted heat seeking missile. During a 90-day pilot test, predictor and criterion data were collected nearly concurrently. Performance data were collected during a Field Training Exercise (FTX) at the Realistic Air Defense Engagement Simulation (RADES) facility using actual weapon systems against subscale models of fixed and rotary wing aircraft.

METHOD

Participants

Data were collected from 26 16P personnel and 75 16S personnel in Advanced Individual Training (AIT) at Fort Bliss. All had received training in visual recognition of aircraft. The 16S trainees had qualified for operation of the Stinger weapon in the Moving Target Simulator. The 16P personnel had been "familiarized" in the operation of the Chaparral weapon system.

Predictors

Psychomotor Skills. Two computerized psychomotor tests were included from the Project A Battery: One-handed Tracking and Two-handed Tracking. The former measures steadiness and precision; the latter measures coordination and dexterity. Scores are in terms of mean distance off target.

Spatial Ability. Two of the Project A paper-and-pencil tests of spatial ability were included. The Maze Test assesses the ability to scan a field. The Orientation Test measures the ability to maintain one's perspective or bearing with respect to some object when it and its component parts have been rotated. Scores on these semi-speeded tests were numbers of items correct.

Other predictors. Predictors also included the official Basic Rifle Marksmanship (BRM) scores obtained for each test participant during Basic Combat Training.

Criterion measures obtained from RADES

Leader/Observer Tasks. The following criterion measures were collected for Leader/Observers on Chaparral and Stinger teams during engagements involving fixed wing (FW) aircraft (i.e., jets): (a) Range at Detection--Range of

¹The views expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Institute or the Department of the Army.

²Now at the Office of Personnel Management

aircraft at detection response in full scale kilometers (full scale distance = measured distance times scaling factor); (b) Range at Identification--Range of aircraft at identification response in full scale kilometers; (c) Time to Identification--Time interval in seconds between detection response and identification response; and (d) Percent Identified Correctly--Number of correct identification judgments ("hostile" or "friendly") divided by number of identification judgments made.

For engagements involving rotary wing (RW) aircraft (i.e., helicopters), measures included: (a) Raw Detection Time--Time in seconds from software command to raise helicopter on stand until Leader/Observer makes detection response; (b) Time to Identification--Same as above; and (c) Percent Identified Correctly--Same as above.

Gunner Tasks. The following criterion measures were collected for Gunners on Chaparral teams during engagements involving FW aircraft: (a) Tracking measures which included: (1) Number of Acquisitions--Number of intermittent acquisitions of aircraft by the weapon system; (2) Total Acquisition Time--Total time in seconds that weapon acquisition of aircraft is maintained; and (3) Percentage of Available Time on Target--Percent of total possible acquisition time window that acquisition is actually maintained; (b) IFF Interrogation--Did Gunner interrogate the aircraft with the Identification Friend or Foe (IFF) subsystem? (c) Time from First Acquisition to Fire--Time in seconds from the first acquisition of the aircraft by the weapon until fire; (d) Total Engagement Time--Time in seconds of the entire engagement from detection response to fire; and (e) Effect or Engagement Efficiency--Number of aircraft hit ("killed") divided by the number of fire events.

Stinger teams were assessed on all of the above criteria with the exception of "Time from First Acquisition to Fire". In its place, Stinger teams performed the following two tasks: (a) Time to Lock-on--Time in seconds between the first acquisition of the aircraft by the weapon heat seeker and press of uncage bar on weapon by Gunner; and (b) Time to Fire--Time interval in seconds between press of uncage bar and fire. During engagements involving RW aircraft, tracking data were not collected; all other listed criteria of Gunner performance were assessed.

RADES Simulation Facility

RADES is a Short Range Air Defense engagement simulation located in the desert at Condon Field, White Sands Missile Range. Soldiers manning actual instrumented weapon systems engage subscale flying FW and pop-up RW aircraft of U.S. and Soviet design. Crew engagement actions are recorded in terms of time and range of aircraft at event occurrence. Time measures are accurate to 250 milliseconds and range measures are accurate to 100 meters. Detailed description of RADES can be found elsewhere (Drewfs, Barber, Johnson, & Frederickson, 1988; Lockhart, Johnson, & Sanders, 1987).

Weapon Systems

Three Chaparral weapon systems (M48A2) were used for each day of 16P testing. Each Chaparral had a Forward Looking Infrared subsystem, firing key, gunner commo headset, MIM72C Tracking Head Trainer (THT) with arming plug, and IFF Subsystem Training Set. Additional required equipment included 3 TA312 field telephones with 3 headsets and 3 spools of commo wire, and 3 pairs of 7 x 50 binoculars. Finally, 1 Chaparral systems mechanic with 24N tool kit was always present during 16P testing. For each day of 16S testing 5 Stinger THTs (M134) were used. Each Stinger THT came with 5 THT batteries and an IFF simulator. Additional required equipment included 5 armor vests and 5 pairs of 7 x 50 binoculars.

Procedure

16P and 16S AIT personnel were tested in RADES during their FTX. Typically, weapons were set-up and calibrated on Friday. Testing took place

all day Saturday and Sunday, and during Monday morning. Monday afternoon the weapons were returned to Fort Bliss.

Upon arrival at RADES the trainees were instructed as to the nature of the testing and what was required of them. Trainees were then assigned to groups for morning or afternoon performance testing. Within each morning or afternoon group trainees were assigned to weapon stations and teams.

Trainees were assigned to weapon positions as two-man teams. Each team consisted of a Leader/Observer and a Gunner. The Leader/Observer was responsible for detecting the aircraft, identifying the aircraft as either friendly or hostile, and issuing a command to engage or cease engagement. The Stinger team Leader/Observer issued his commands to the Gunner directly, since he was standing next to him at the weapon position. The Chaparral Leader/Observer issued his commands to the Gunner via the field telephone, since the Gunner was positioned inside the closed turret of his weapon 5 to 10 meters distant. The Gunner was responsible for interrogating the aircraft, acquiring the aircraft with the weapon, tracking the aircraft, ranging the aircraft, locking-onto the aircraft, superelevating and leading (Stinger only), and firing.

At a weapon position, the data collector reviewed the engagement actions with the team and showed them their sector of responsibility and primary target line. Each team was responsible for defending the same 90 degree sector of sky. All five weapon positions were visually and aurally independent of one another and no cross cuing was permitted. Each team was instructed in the discrete trial procedure employed, and reminded of the trial begin and trial end signals. Each trial began when the data collector gave the team an alerting and cuing message. This message stated that air activity was imminent, reminded the team that their Weapons Control Status was "tight", gave a clock azimuth cue to the predicted aircraft ingress avenue, and identified the intruder as either high or low in elevation. A typical alert was "red, tight, 1 o'clock, low!". The Primary Target Line was always 12 o'clock. The data collector signalled the end of a trial by alerting the team that air attack was unlikely ("return to status white").

Each team received a total of 14 data trials, 1 trial each on 14 different scenarios. These 14 scenarios are described below by set. Scenario sets A and B each contain seven scenarios. Three are hostile FW, two are hostile RW, and two are friendly RW. Each team completed both sets. A team performed one set with one team member as Gunner and one as Leader/Observer then switched duty positions and performed the second set.

Table 1: Test Scenarios

Scen	Type	Intent	Target	Azimuth	Aspect	Range	Duration
Set A							
1	FW	H	MIG-27	12:00	0 deg		
2	FW	H	Su-17	1:00	45 deg		
3	FW	H	Su-25	11:00	45 deg		
4	RW	F	AH-1	11:00	0 deg	3 km	25 sec
5	RW	F	CH-3	1:00	45 deg	5 km	25 sec
6	RW	H	M1-24	12:00	45 deg	3 km	25 sec
7	RW	H	M1-8	11:00	0 deg	5 km	25 sec
Set B							
11	FW	H	MIG-27	1:00	45 deg		
12	FW	H	Su-17	11:00	45 deg		
13	FW	H	Su-25	12:00	0 deg		
14	RW	F	AH-1	11:00	0 deg	3 km	25 sec
15	RW	F	CH-3	1:00	45 deg	3 km	25 sec
16	RW	H	M1-24	12:00	0 deg	5 km	25 sec
17	RW	H	M1-8	11:00	45 deg	5 km	25 sec

Hypotheses

We hypothesized that psychomotor test scores would predict Gunners' tracking performance for FW engagements and spatial test scores would predict Leader/Observers' RW detection times. Spatial and psychomotor test scores were not expected to directly predict engagement efficiency since these predictors do not assess the requisite skills. However, better tracking performance could result in higher engagement efficiency and earlier RW detection could result in lower percent of hostile aircraft delivering ordnance.

RESULTS

Table 2 displays means and standard deviations for predictors and shows correlations between predictors and criteria.

Findings for 16P. None of the hypothesized relationships emerged. With few exceptions, relationships between predictors and criteria were not significantly different from zero. These exceptions included significant correlations between spatial test scores and the number of intermittent acquisitions ($r = .47$, $p < .05$), and between BRM scores and time taken to identify the target ($r = .52$, $p < .05$). Both correlations were opposite from the expected direction. Thus, soldiers with higher psychomotor test scores tended to wait longer to identify targets and lose IR acquisition of the target more often.

Findings for 16S. Again, none of the hypothesized relationships emerged. However, soldiers with higher BRM scores tended to detect targets at greater distances ($r = .40$, $p < .05$). Further, soldiers who have higher psychomotor and spatial test scores tended to track for longer periods of time and to wait longer to identify targets. Tracking ability correlated negatively with the range at which Leader/Observers identified the target in FW trials ($r = -.32$, $p < .01$). Soldiers with more spatial ability tended to wait longer to identify both jets ($r = .35$, $p < .01$) and helicopters ($r = .28$, $p < .05$). In the Gunner role, soldiers scoring better on spatial tests tended to track jets longer ($r = .31$, $p < .01$), but tended complete RW engagements more rapidly ($r = -.25$, $p < .05$).

DISCUSSION

In summary, none of the hypothesized relationships emerged. As suggested by past research, (Mikos, Casey, & Lockhart, 1980), the specific spatial and psychomotor abilities measured during this evaluation may in fact, be unrelated to short range air defense skills. However, a number of measurement issues, outlined below, prevent us from drawing this conclusion at this time. These issues preclude a fair test of the Project A predictors.

First, any possible relationships may have been obscured by the small sample size, especially in the Chaparral sample ($N = 19$ for FW; $N = 26$ for RW). With an N of 19, the 95% confidence intervals around the observed correlations is $\pm .45$. For the Stinger sample ($N = 68$ for FW; $N = 75$ for RW) these confidence intervals were $\pm .24$.

Second, many Gunners in both occupations failed to perform key steps. Thus, when data were aggregated across trials, the aggregated scores were based on only those trials for which data was available. As a result, composite scores may have underestimated true differences in individual performance, thus minimizing any relationships among predictors and criteria.

Another limitation was that in many instances, we were attempting to predict team performance from individual attributes. The RADES testbed was designed to measure team (unit) performance. The Project A predictors were designed to predict individual performance based on individual attributes. For many of the criterion variables Leader/Observers' and Gunners' performance cannot be separated. For example, the total time available for a Gunner to track a target depends greatly on how early in a trial the Leader/Observer

Table 2

Means and Standard Deviations of Criterion Measures and their Relationships with Predictors for
16P Personnel

		BRM ^a	PsMtr.	Spat.	P/S Comp.	
Mean		28.78	55.77	57.18	162.73	
S.D.		3.98	7.95	10.07	21.25	
	MEAN	SD	Correlations			
<u>Fixed Wing (Jet) Trials</u>						
Range at Detection	9.50	3.27	.38	.10	.08	.12
Range at Identification	3.88	2.23	-.10	.08	.18	.15
Time to Identify	23.54	10.97	.52*	.07	-.12	-.01
% Identified Correctly	84.21	25.19	.03	-.02	-.34	-.19
<u>Rotary Wing (Helicopter) Trials</u>						
Raw Detection Time	27.03	3.02	.18	.08	-.12	.00
Time to Identify	10.37	4.03	.04	.03	-.11	-.04
% Identified Correctly	71.50	22.66	.10	.03	-.20	-.07

GUNNER TASKS

<u>Fixed Wing (Jet) Trials</u>						
No. Acquisitions	4.11	2.58	.08	.38	.47*	.55*
Tot. Acquisition Time	10.74	7.74	-.27	.08	.03	.07
% Available Time on Target	49.32	23.58	-.20	.13	-.37	-.10
Time from First Acquisition to Fire	21.32	9.56	.25	.01	.11	.07
Total Engagement Time	38.96	12.71	.44	.08	.29	.22
Effect	71.79	31.80	.21	-.31	.11	-.18
<u>Rotary Wing (Helicopter) Trials</u>						
Time from First Acquisition to Fire	13.19	4.50	.03	.00	-.03	-.02
Total Engagement Time	16.21	5.13	.07	.15	.28	.25
Effect	31.65	34.94	-.13	-.07	-.13	.13

LEADER/OBSERVER TASKS

COMP.

<u>FW Trials</u>						
Range at Detection	9.21	2.21	.40*	-.04	.15	.03
Range at Identification	4.06	1.64	.13	-.32**	-.22	-.33**
Time to Identify	21.05	8.89	.24	.26*	.35**	.33**
% Identified Correctly	88.29	21.13	-.15	-.27*	-.10	-.24*
<u>RW Trials</u>						
Raw Detection Time	25.45	2.66	.02	-.03	-.13	-.08
Time to Identify	10.72	3.85	-.17	.07	.28*	.16
% Identified Correctly	67.29	21.04	.04	-.15	.08	-.07

GUNNER TASKS

<u>FW Trials</u>						
No. Acquisitions	2.40	1.77	.06	.08	.14	.11
Tot. Acquisition Time	6.54	4.48	.07	.12	.31**	.20
% Available Time on Target	81.47	21.95	.03	-.15	-.02	-.12
Time to Lock on	9.28	6.24	.10	.14	.23	.19
Time to Fire	3.15	2.50	-.14	-.04	-.17	-.10
Total Engagement Time	28.88	8.73	-.30	.09	-.06	.04
Effect	87.28	26.09	-.03	-.26*	-.06	-.21
<u>RW Trials</u>						
Time to Lock on	6.78	3.45	-.01	-.18	-.04	-.15
Time to Fire	4.03	2.42	-.21	.00	-.16	-.06
Total Engagement Time	15.90	5.43	-.32	-.13	-.25*	-.19
Effect	41.07	35.75	.28	.06	.16	.11

Note. One asterisk (*) indicates a significance level of $p < .05$; two asterisks indicate a significance level of $p < .01$.

^aN = 35 for FW and 36 for RW Trials for BRM; N = 68 for FW and 75 for RW for Psychomotor, Spatial, and P/S Composite.

detects the target and how late he issues the visual identification and engagement command.

Finally, attributes other than those measured by this set of predictor tests may be more effective predictors of performance on missile based weapons systems. In this pilot, we used tests that had been used at other sites to predict gunner performance (Smith & Walker, 1988, Smith & Graham, 1988). Relationships among the RADES criterion measures suggest that Gunners who perform their requisite actions more quickly have greater engagement efficiency. Thus, measures of perceptual speed and accuracy included within the Project A Computerized Predictor Battery might be better predictors of Short Range Air Defense skills.

REFERENCES

- Barber, A.V., Drewfs, P.R., & Johnson, D.M. (1987). Performance of Stinger teams using the RADES multiple weapon configuration. (Working Paper FB 87-09). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Barber, A.V., Drewfs, P.R., & Lockhart, J.M. (1987). Effective Stinger training in RADES. (Working Paper FB 87-02). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Drewfs, P.R., Barber, A.V., Johnson, D.M., & Frederickson, E.W. (1988). Validation of the Realistic Air Defense Engagement System (RADES). (ARI Technical Report 789). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Eaton, N.K., Goer, M.H., Harris, J.H., & Zook, O.M. (1984). Improving the Selection Classification and Utilization of Army Enlisted Personnel: Annual Report, 1984 Fiscal Year. (Technical Report No. 660). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Johnson, D. M., Barber, A. V., & Lockhart, J. M. (in press). The effect of target background and aspect angle on performance of Stinger teams in the Realistic Air Defense Engagement System (RADES). (ARI Technical Report). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Lockhart, J.M., Johnson, D.M., & Sanders, W.R. (1987). Manpower, personnel and training analysis of air defense crews in RADES. Proceedings of the 26th Annual U.S. Army Operations Research Symposium, pp. 221-231.
- Mikos, R., Casey, R.J., & Lockhart, J.M. (1980). Validation of research on acquisition and retention of cognitive versus perceptually oriented training materials. (ARI Technical Report 514). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Nunnally, J.C. (1978). Psychometric Theory. (2d ed.). NY, NY: McGraw-Hill.
- Smith, E.P. & Graham, S.E. (1987). Validation of psychomotor and perceptual predictors of Armor Officer M-1 gunnery performance. (ARI Technical Report 766). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.
- Smith, E.P. & Walker, M. (November, 1988). Testing psychomotor and spatial abilities. Paper presented at the 30th Annual Conference of the Military Testing Association.

MAINTAINING COURSE CURRENCY IN A CHANGING ENVIRONMENT

Mollie J. Iucker

U. S. Naval Guided Missiles School
Dam Neck, Virginia 23461-5250

ABSTRACT

During the next decade, emerging technologies in education and training will continue to affect Navy training. Rapidly changing technologies will make heavier demands on the information processing and decision making capabilities of training facilities responsible for initiating and maintaining control of Navy curricula. Curricula changes will require integrated responses making communication not only essential but critical.

This paper describes the procedures for changing curricula at the Naval Guided Missiles School (NAVGMSCOL). The model demonstrates the procedures for maintaining control of curricula that must be responsive to changing training requirements and equipment/documentation alterations. The process is described from the initial identification of a curriculum deficiency to the distribution of the change material to the user. Maintaining course currency through a surveillance and change program is an integral part of the Navy's response to a changing environment. By following the steps in this model, timely responsiveness in such a highly technical complex environment can be a reality and not a dream.

BACKGROUND

Historically, the U. S. Navy's Strategic Weapon System (SWS) Training Program has been viewed as dynamic and responsive. Changes are constantly made to keep pace with technological advances in hardware, operating techniques and training requirements. To meet the goal of the SWS Training Program, Naval Guided Missiles School provides training to ensure the SWS can be operated and maintained in a high state of readiness.

Formal training is presented by naval instructors using Instructor Guides supported by Trainee Guides, technical documentation, instructional media materials, laboratory equipment, laboratory exercises and tests. Training materials are designed, developed and implemented to accurately reflect changes in the training requirements. A number of stimuli can provoke a change in the training materials. However, regardless of the reason for a change, the procedures to ensure that the training materials are accurate and current remain static. The Instructor Guide is the principal element of every curriculum and provides basic course programming. When promulgated by the Chief of Naval Technical Training (CNTECHTRA), the Instructor Guide is the master plan for instruction. When the same course is taught at another training facility (TF), each facility must teach the approved curriculum. Modifications to curriculum materials are documented in the Instructor Guide.

CURRICULUM AUTHORITY

A Course Curriculum Model Manager (CCMM) is designated for each SWS course. NAVGMSCOL is the CCMM for 34 SWS courses. Operating under a functional commander who has curriculum control authority, NAVGMSCOL is responsible for surveillance and maintenance of these courses. As the CCMM, NAVGMSCOL is also responsible for initiating curriculum development and revision, conducting course reviews, maintaining course audit documentation and budgeting for production and distribution of curriculum materials. The Curriculum and Instructional Standards Office (CISO) serves as the quality control processing link for all curriculum changes.

SWS training materials remain current and accurate through surveillance and change efforts. Curricula surveillance requires the communication and cooperation of all SWS personnel. Quality control is the shared responsibility of each member of the SWS community.

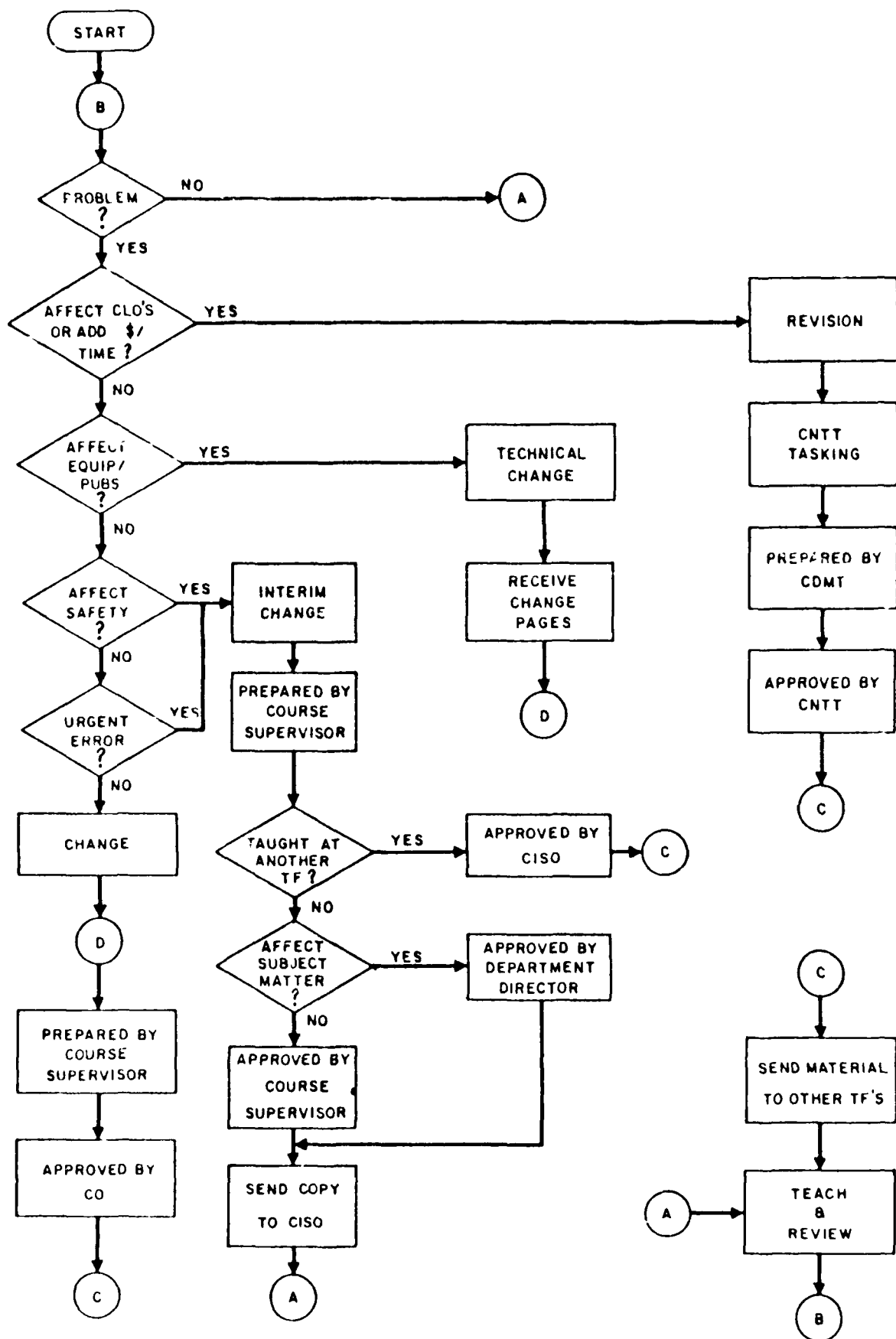
Modifications to existing curriculum are accomplished through a revision, technical change, change or interim change process. The method of change is determined by the magnitude and urgency of the change. Following higher authority directives, Figure 1 summarizes the process used by NAVGMSCOL as CCMM for identifying and implementing modifications to the curriculum. The questions that must be answered are identified by decision blocks. The procedures are identified by process blocks. The teach and review surveillance process occurs continually and is identified in the figure by the A and B connectors.

COURSE REVISIONS

NAVGMSCOL is responsible for requesting a curriculum revision for those courses when it has been designated CCMM. Other training facilities can recommend that a course be revised, but only the CCMM can officially request permission to revise a course. Course revisions are necessary when one of three conditions exist: (1) The curriculum requires a change to the course learning objectives (CLO's), (2) there is an increase in the course length or (3) additional resources are needed. When a revision is justified, the revision recommendation is approved by CNTECHTRA. CNTECHTRA may task NAVGMSCOL or a contractor to develop the revision. If NAVGMSCOL receives the tasking following development by the Curriculum Development and Maintenance Team (CDMT), the revision is approved by CNTECHTRA. The Curriculum and Instructional Standards Office (CISO) distributes the revised materials to other training facilities. The distribution of materials to other training facilities is shown in figure 1 by the C connector. Once the revised materials are promulgated and implemented, the teach - review surveillance effort begins.

TECHNICAL CHANGE

Changes to tactical and training unique equipment or publications will result in a technical change. Technical changes are usually forwarded to NAVGMSCOL by the contractor responsible for providing training support and/or equipment surveillance. Upon receipt, the technical change is reviewed by the chain of command and promulgated as a change. The continuation of the process is shown in figure 1 by the D connector.



NAVGMSCOL CURRICULUM CHANGE MODEL

Figure 1

CHANGE

When the three conditions requiring a revision do not exist, training materials are modified by a change process. Minor and urgent changes are incorporated through a special type of change called the interim change. This is described under a separate heading. Major modifications to the topic learning objectives, discussion points, other elements in the instructor guide or trainee guide and instructional media materials are accomplished through the curriculum change process. Inserting updated topical information, shifting instructional strategies and resequencing topics in the course are usually executed by a change.

Changes are prepared by the course supervisor, reviewed by the chain of command and approved by the Commanding Officer (CO). If the course is taught at other training facilities, copies of all changes are distributed to those facilities by the CISO.

INTERIM CHANGE

An Interim Change (IC) is the quickest way to alter training materials. This is considered the quick fix method. This is a minor change that is an urgent response to an identified deficiency in a course. Unlike the revision, technical change or change, any training facility can correct a minor deficiency in the training materials using the IC. IC's are used to change or update reference publications, insert Type Commander promulgated subjects, increase teachability in the topics of instruction, correct typographical or technical errors and eliminate unsafe training practices. IC's are temporary pen and ink changes and are incorporated into the next formal curriculum change or revision.

When the course is only taught at NAVGMSCOL, the Course Supervisor prepares and approves the IC. If the IC involves a subject matter change, it is routed through the departmental chain of command for approval by the Department Director. A copy of all approved IC's are sent to CISO for filing in the master course file.

If the course is taught at other training facilities, and NAVGMSCOL is the CCMM, the IC is prepared by the Course Supervisor and routed via the chain of command to CISO. CISO approves and distributes the IC to other training facilities teaching the course. The teach and review process continues.

CONCLUSIONS

The model shown in figure 1 is based on a systems approach to curriculum change. The procedures bring together the subject matter experts and the quality control experts. Communication is essential at each step. The constant monitoring of the hardware and documentation changes and the uninterrupted review of curriculum materials involve correcting errors and deficiencies as they are detected. Through this process, the effectiveness of NAVGMSCOL in meeting the needs of the SWS program is ensured.

REFERENCE

Chief of Naval Technical Training (1987, May) Single Standard Guidance. Millington, TN.

Navy Classroom Training: A Status Report

Barbara Taylor

John Ellis

Robyn Baldwin

Navy Personnel Research and Development Center
San Diego, CA 92152

Introduction

The Navy currently teaches over 7000 courses. The great majority of these courses are presented in traditional classrooms and laboratories with a group of students taught by a single instructor. Almost all of these courses were developed using some version of the systems approach to training development. Recently, the Navy has adopted a single standard for instructional development, MILSTD-1379C and DOD-HDBK-292: Training Materials Development. Prior to the adoption of this standard, each branch of the Navy had its own set of guidelines for developing instruction.

The Navy has used a systems approach to training development (in a variety of forms) for over two decades. When properly applied, this approach results in training that has a number of desirable qualities. These include: (1) a detailed analysis of job requirements, (2) job relevant learning objectives, (3) an instructional program designed to assure that students achieve the the objectives, and (4) tests that verify this achievement.

We recently conducted an extensive evaluation of Navy classroom training to determine how effectively the systems approach was being applied in the three Navy warfare communities (air, surface and subsurface).¹ We evaluated 100 enlisted Navy courses. The courses were selected to be representative of the total number of enlisted training courses by warfare community and by type of course (A , C, and F Schools). Since the focus of our evaluation was on classroom training instead of laboratory instruction, we concentrated on knowledge rather than skill objectives. For each course we reviewed a representative sample of the course objectives, test items, and instructional presentation. This usually amounted to at least one week of instruction and may have been more than one week for lengthy A and C schools. Note than none of the courses we evaluated were developed under the new single standard, although several were developed under the submarine guidelines which are the basis for the new standard.

Results for Course Objectives

The first major finding was that 56 percent of the objectives did not match the course training goal identified by the course instructors. A training goal is the target performance level for job entry. There are three general training goals to which Navy technical courses train: heavily supervised on-the-job training (OJT), minimally supervised on-the-job training, and skilled performance. For heavily supervised OJT, the student should be generally familiar with terminology, technical documentation and duties required on the job. For minimally supervised OJT, the student should be able to

perform the job with very little supervision. For skilled performance, the student should be able to perform the task with no assistance from a supervisor. If any part of an objective did not conform to training goal the entire objective was judged inappropriate.

There were differences among warfare communities in the percent of inappropriate objectives. The air community had the most (65 percent) and the submarine community the least (39 percent). The surface community was close to the submarine community (44 percent). One explanation for this difference is that the air community identified the training goal for a number of courses as minimally supervised OJT. This level of training requires that the passing score for the courses be relatively high. The air community passing scores, specified in their objectives, was either 63 or 75 percent for all the objectives we evaluated. This is too low a passing score for minimally supervised OJT. The submarine community, on the other hand, identified heavily supervised OJT as the training goal for the majority of the courses we evaluated. For this training goal an objective that requires a 75 percent passing score is considered appropriate.

As part of our evaluation of objective appropriateness, we verified the judgements of course instructors on course training goals by conducting a survey of supervisors of recent graduates. For 61 of the 100 courses, we asked one or more supervisors at what level the graduates were when they arrived and at what level they needed the graduates to be upon arrival. For all courses with a heavily supervised training goal as identified by the school, the majority of the fleet supervisors contacted said that was their requirement and what they received. For over 35% of these courses, the fleet contact needed and received a graduate able to perform at the minimally supervised level. In other words, for over one-third of the courses taught to the heavily supervised level, the fleet contact felt that they received a graduate able to perform at a higher level. For the courses that the school identified as having a minimally supervised training goal, the vast majority (85.72%) of the fleet contacts felt that they needed and received a graduate able to perform at that level. The fleet contacts reported that for 80% of all courses with a school identified training goal of skilled performance, they received a graduate able to perform at that level. For the remaining 20%, the fleet contacts stated that they needed and received a graduate able to perform at the minimally supervised level. Again, they got what they wanted.

Even though the majority of the objectives were judged inappropriate, it is encouraging that in almost all cases the schools identified a training goal at the level required by the fleet, and the fleet personnel we questioned said that graduates were able to perform at that level. One explanation for this discrepancy between objective appropriateness and student job performance is that many of the instructors we interviewed said that even though the minimum passing scores were only 75 percent, they worked with students to attain higher levels of achievement. These comments by instructors are supported by the fact that average passing scores for many of the courses we evaluated exceeded 85 percent. The instructors seemed to realize that the standards mandated by the objectives were insufficient for reliable job performance.

Results for Test Items

The major finding was that half of the objectives were not tested. This happened in spite of the fact that 92 percent of the objectives were rated essential for job performance

by the course instructors. There were again differences among the warfare communities in percent of objectives not tested with the air community having the most (58 percent) and the surface and subsurface having substantially fewer (34 percent and 27 percent, respectively). A possible reason for the air community failure to test objectives is the large number of very specific objectives that were developed for each lesson topic. It was not uncommon for a one hour lesson to have over a dozen supporting objectives. A weekly 500 question test would have to be given to cover all these objectives.

An additional finding was that for objectives that were tested, 48 percent of the test items did not match the requirements specified in the objective. Once again there were differences among communities. In the submarine community 93 percent of the test items did not match what the objective required, while only 44 percent of air and 39 percent of surface community test items failed to match. The primary reason for this difference is that the submarine curriculum development guidelines do not require that the objective and test item behavior be written at the same level. For example, it is acceptable to have an objective that requires the student to be able to recall information from memory and a corresponding test item that only requires the student to recognize information by selecting the correct answer from a list of distractors. An objective requiring recall might also be tested by having the student look up the information in a technical manual.

The failure to test objectives and to write test items that match objectives pose potential problems even though the fleet survey showed that the fleet was getting what it expected. The fleet survey was probably not sensitive to these problems because the fleet personnel evaluated graduates on job skills. The objectives we examined were concerned with job knowledge not job skills. However, problems with job knowledge can affect on both on-the-job training and classroom/lab performance. Deficiencies in knowledge objectives can result in inefficient instruction, more time spent in shipboard on-the-job training to reach proficiency, and training that is academic instead of job oriented. These problems would not be obvious to the fleet personnel we surveyed because they would not be aware that the quality of job knowledge could be improved and training could be made more efficient and job oriented.

We found an example of the kind of problem that can result from inattention to knowledge objectives in our own data. We divided knowledge objectives into two categories; those that were tested with open book tests and those that were tested with closed book tests. We then compared the frequency of each type of objective across the three warfare communities. The reason we did this was to see how job oriented each community was with respect to knowledge objectives. In technical training most knowledge objectives should be tested using open book tests because the training should emphasize the use of technical manuals/documentation vice memorization. The exceptions to this are knowledge objectives concerning safety requirements that need to be memorized and objectives dealing with background information that must be known to be able to use the technical manuals. Although these exceptions are important, they still should represent a minority of knowledge objectives. We found that over 70% of all objectives within the surface and the submarine communities were tested with closed book tests. In the air community, however, the majority of the objectives were tested with open book tests. The air community emphasized the use of technical manuals and students were taught how to locate information rather than memorize it. This is important

because memorized information is often forgotten especially if it is not used frequently. A locate type of objective is also desirable because it involves students in activities similar to those that occur on the job. In summary, the failure to develop and test appropriate job-oriented knowledge objectives can cause problems both inside and outside the classroom that could not be easily detected by fleet personnel. This is especially true for A schools which handle large numbers of students annually and have a high concentration of knowledge objectives. Fleet expectations for graduates of these schools are not very high (all the A schools were rated heavily supervised OJT by fleet personnel) and they may not realize the product could be improved.

Results for Instructional Presentations

The primary finding was that the instructional presentations contained all the information required by the learning objectives and were effectively presented by the instructors. There were two important exceptions to this finding. The first was that there was a significant lack of opportunities to practice in submarine community courses. This was true both in lectures and in the student handouts (e.g. workbooks, note taking sheets, etc.). Since practice has shown to greatly increase retention, and is one of the most efficient uses of instructional time, this is not a trivial problem.

The second exception was that many instructional practices and innovative technologies that have shown to have a positive effect on student learning in civilian classrooms are not being utilized within military training. Some examples of effective instructional practices include peer instruction (sometimes called cooperative learning), advance organizers, external rewards, innovative remediation strategies, and systematic assessment of student progress. In regard to innovative technology, the classroom we typically saw during our evaluations could have been a classroom 25 years ago. Media almost always consisted of a blackboard and occasionally a few transparencies. The transparencies were usually copies of tables or diagrams from technical manuals. In short, the benefits of training technology have not reached the typical Navy classroom.

Conclusions.

The problems that we found are primarily due to objectives and tests that were written without regard to the level of knowledge required to support future job requirements. This accounts for the rather overwhelming number of objectives and test items that we classified as inappropriate. This is not to say that the majority of graduates will not be able to perform at the level expected of them. On the contrary, our fleet survey indicated that fleet supervisors identified a training goal consistent with the graduates they received. Instructors that we interviewed often stated that the test passing scores were too low and that they would work with students in order to attain higher test scores. However, it is important to note that students that just meet the minimum requirements do pass the course. This problem is not a difficult one to fix. The notion of training goals and related test passing scores is not a foreign one to those who develop instruction. The Navy's instructional development guidelines should eliminate the notion of global passing scores that are applied across large numbers of courses and instead let passing scores be set for each individual course based on job requirements.

The problems with untested and/or tests mismatched to objectives are also not difficult to solve. The instructional development guidelines should be revised to promote the development of more job related objectives and test items. There should be less emphasis on rote memorization and more emphasis on comprehension and familiarization with the technical documentation. They should provide instructions on how to write objectives and test items to insure that all objectives get tested and that the test items are written so they match the objectives.

It is important emphasize that the Chief of Naval Education and Training is aware of these problems and the the Navy is in the process of revising the new single standard and associated instructions to correct them.

One important area in which there were no problems was the instructional presentation. The Navy's instructor training and the subsequent certification process (accomplished in the schools) is producing highly qualified instructors. The instruction we observed was well presented, the instructors were enthusiastic, and students were attentive. The only deficiency in the classrooms is the lack of modern media. Most instructors used the blackboard (or white board) and transparencies containing information as it appeared in a technical manual. Many of the presentations could have been enhanced with additional media focused on the instructional objectives. As all types of media become more and more affordable, efforts should be made to explore upgrading Navy classroom presentations.

In addition, some instructional practices that have been shown to have a positive effect on learning within the civilian community such as peer instruction and systematic assessment of student progress were not utilized in the schools we observed. These practices should be applied in Navy schools on a trial basis to see how well they work.

Finally, although the Navy's new instructional development guidelines are undergoing revisions and emerging instructional techniques and technologies need to find their way into Navy classrooms, it is should be emphasized that most Navy courses do not need major revisions. In fact the majority of courses would realize little benefit from attempts at improvement primarily because over 70 percent of Navy courses have class sizes less than ten and many of them have five or fewer students. With classes this small the individual instructors can usually overcome problems inherent in the instructional objectives, tests, presentation and media. In small classes it is easy for an instructor to interact with individual students and to monitor student progress. A small class just naturally lends itself to that type of activity. Furthermore, courses with small classes are frequently equipment oriented. In this type of course students spend most of their time working in laboratories rather than sitting in classrooms. Therefore, we recommend that future research and attempts to enhance classroom training should concentrate on courses which have large classes (15 or more), a large annual throughput, and have a significant amount of instructional time invested in classroom training. Improvements in these high through-put courses will save the Navy training dollars and improve readiness by shortening the amount of time it takes an apprentice to become a journeyman.

¹ For a more detailed accounting of this project see NPRDC TR-87-19, The Course Evaluation System, and NPRDC TR-88-11, The Current Status of Navy Classroom Training: A Review of 100 Navy Courses with Recommendations for the Future

Overseas Mail Time for Correspondence Instruction

Grover E. Diehl
USAF Extension Course Institute

This study was the seventh in a series of reports evaluating the transit time and delivery condition of Extension Course Institute (ECI) courses. In most respects, it was a replication of the sixth report in the series and included a complete summary of the preceding research.

After reviewing the evidence available in 1983, it appeared that bona fide mail delays were very few--both in number and in location. Unfortunately, identification of the highly unusual event was perhaps the most important limitation of the preceding studies. Kelly (1976) and York (1980) used aggregated data; York and Diehl (1981) sampled too few sites. Information from the Military Postal Service Agency (MPSA) and US Postal Service (USPS) contained sufficient detail, but the procedures for measuring elapsed time missed periods which may have contributed significant delays, one of these being local distribution. To control for these and other factors, the sixth study was conducted which surveyed a larger sample, increased the number of sorting categories, expanded the response set on the questionnaire, used a considerably more sophisticated data analysis process, and increased the accuracy of data capture. This study, How Long Does It Take ECI Courses to Reach Students? (Diehl, 1984), sampled 8000 students worldwide and had a return rate of 35.75 percent. There were five major conclusions:

1. Fifty percent of students received their courses in under 19 days, 80 percent in under 28 days, 90 percent in under 35 days, and 95 percent in under 42 days. (These figures were uncorrected and may have included time during which students were on leave, TDY, or otherwise unavailable to receive mail.)

2. Ninety-eight percent of the mail was received in fully satisfactory condition.

3. About two percent of students complained about long delays.

4. Professional Military Education (PME) students were six times more likely to complain about mail delays than either Specialized or Career Development Course (CDC) students, although transit times for all three groups were the same. PME students were nearly three times more likely to make fatal mistakes in filling out the survey.

5. With the exception of the Panama Canal Zone (APO Miami), differences among locations, countries, and gateway post offices were either not demonstrated or inconclusive. Canal Zone mail ran seven to ten days faster than other overseas mail. Installations with high volumes of mail and APO mail were slightly faster than low volume locations and FPO mail.

Unfortunately, ECI continued to receive mail time complaints, so an update of the 1984 results was in order. In addition, there were several upcoming operational changes supporting a new mail survey. For one thing, the supply and warehousing facility changed contractors in November 1987, which made timely, accurate data especially important. There were also changes in the mailing labels which would make future surveys of this type particularly complex and costly.

Findings

Survey Return Rates. In all, 2,753 surveys of 3,049 were returned, for a response rate of 30.42 percent. This rate was about five percent lower than

the 1984 study, which had a return rate of 35.75 percent. The number of usable returns, 2,554, was about 92.77 percent of total returns, also down slightly from 1984 (94%). The mix of returns was almost identical to the mix of the mailings.

The largest dislocation in returns was in the rejected surveys. After correcting for sample size, PME students were almost three times more likely to make fatal errors in filling out the postcards (12.4%) than either Specialized (4.7%) or CDC (3.6%) courses. Virtually all of the PME errors were in receipt date. This finding paralleled the 1984 study (a multiple of 2.8), where it was noted that the finding was perplexing since PME courses contained a relatively higher proportion of officers and senior enlisted students than either Specialized Courses or CDCs.

Also corresponding to the 1984 data, the proportion of PME, Specialized and CDC courses going overseas was slightly different from total enrollments, the difference being a slightly lower proportion of PME students. No reason for this finding has been found, and it remains possible that overseas enrollments comprise a slightly different subset of students than those in CONUS.

Overall Transit Time. The mean delivery time was 20.512 days (standard deviation (SD) of 11.296 days). The 50 and 80 percentiles were 17 and 25 days, respectively. These times were slightly improved from 1984 where the transit times were 21.745, 19, and 28, respectively. There was a 90 percent delivery rate by day 33 and 95 percent delivery rate by day 43. The range was two to 98 days and the distribution was positively skewed (third moment = 2.540).

An apparent ambiguity in the data involved transit times of five days or less -- 14 cases. While it was shown in 1984 that some locations routinely received course packages in four days (the Panama Canal Zone was cited), a delivery time of two and three days would be considered suspect except for the fact that the data were reported by the recipients. While transit times of two and three days may be unlikely, they seem to be facts to the students.

Equally perplexing were transit times in excess of five weeks (35 days), mainly due to the lack of student mail time complaints in this region. Out of 232 courses which were reported to take 35 days or longer, only 13 (5.6%) complained. There were a number of reasons for this outcome, all reported by the students. Many volunteered that they did not receive any mail for extended periods due to leave, TDY, or transit to another location. In these cases, the "receipt dates" were best guesses or the date the students actually received the packages in hand and not when the packages were placed at their mail drop locations. In either case, the reported dates are inflated. Also, having already incurred a delay in receiving all their mail, these students were somewhat less likely to open the CDC packages immediately, further postponing receipt. Given these caveats, "true" delivery times in excess of five weeks were probably not more than five percent of the total (as compared to 9% obtained directly from the raw data).

Type of Course. It was demonstrated in 1984 that PME students were more likely to complain about mail time than other students. There was, then, an interest in determining whether PME courses were delayed here as well. Data showed the USPS and MPSA mail systems continued to be blind to type of course. PME courses averaged 21 days, with Specialized Courses and CDCs averaging 19.5 and 20.4 days, respectively. The small differences noted were not statistically significant, an unusual outcome given the extreme sensitivity of the F-test with 2 and 2551 degrees of freedom. Also, the 50th and 80th percentiles were quite similar.

Location. The quickest average transit time was obtained by San Miguel in

the Philippines, 11.3 days. Although Braisila and Cairo had faster times, 10 and 11 days, respectively, both returned only one survey each. Among the large installations, Howard had the fastest time at 15.4 days. The finding for Howard was slightly longer than in 1984, which was 15.4 days; in 1984 Howard was also the fastest major installation. A number of individual times were considerably faster than these figures; averages are used here, however, to reduce the effect of reporting error.

The longest mail time was at Offenbach, 59 days, followed by Copenhagen, and Hofn, 52 and 51 days, respectively. All three locations returned only one survey. Buchel and High Wycombe returned three or four surveys and had times averaging 51 and 50.5 days, respectively. Among the large installations, Croughton, with 38.1 days, and Preum, with 35.3 days, had the longest times. Generally, though, as volume of mail increased, the average time approached the mean of all returns.

The F-test for locations was significant ($F(152,2399)=3.5959$, $p<.0000$), as expected given the sensitivity of F with a sample of this size. Of more interest was the eta-squared value, .1856, indicating that about 18 to 19 percent of the variance in delivery time was accounted for by location. Unfortunately, post hoc contrasting was not possible due to the number of independent variables. Thus, the pattern of differences remained undefined.

Geopolitical Unit. Omitting geopolitical units (GU) with nine or fewer returns, Panama was the fastest, averaging 15.6 days, and Hawaii was the longest with a 25.9 day average. The fastest time overall was Egypt, 11 days, and the longest overall was Denmark, 52 days; both GUs returned only one survey each. Among seven GUs handling the largest volumes of mail (about 79% of all responses), the Philippines, closely followed by Korea, were the fastest with averages of 17.0 and 17.4 days, respectively. Interestingly, half of the mail to both of these GUs took only 15 days to arrive.

The F-test on GU was significant ($F(29,2524)=4.6381$, $p<.0000$), although the eta-squared was only moderate, .0533; about five percent of mail time variability being associated with geopolitical unit. Post hoc contrasting did reveal a finding of interest: Hawaii was significantly different from Korea and the Philippines. All three were in the same geographic region, but Hawaii was predominately serviced by the USPS while the other two were served by the MPSA. There were no other significant contrasts identified for GUs.

Volume of Mail. Low flow locations took about three days longer than high flow (in 1984 the difference was about four days longer for the low flow locations). While volume of mail was statistically significant, the impact had no practical importance (eta-squared indicating that only slightly over 1% of the variability of mail time was due to volume).

Type of Mail. FPO was fastest, averaging 17.9 days and USPS slowest averaging 23.4 days. APO mail was 20.1 days. The FPO finding was consistent with data presented above pertaining to Pacific mail times; CDCs going via the FPO system are predominantly in the Pacific area. Also, Hawaii, one of the longest mail times, was dominated by the USPS. The finding on FPO mail contrasted with 1984 data, which found APO mail slightly faster.

The F statistic for this ANOVA was of particular interest. The outcome was significant ($p<.0000$), although the eta-squared value was under one percent (.0098). A significant post hoc contrast, however, indicated that APO and FPO were both significantly different from USPS. This was not be taken as an indictment of the USPS generally, but merely recognition that CDC packages going to USPS overseas locations took three to five days longer on average than mail transiting via the MPSA. Overall, though, the size of the effect was very small in the overall computation of mail times.

Geographic Region. An analysis by geographic region (GR) was an attempt to examine mail transit time from a direction other than that used in the normal MPSA, USPS and Air Force channels. Transit times for 12 GRs were determined. Fastest mail time was for Africa, 11 days, which, in fact, was a single course going to Cairo. Latin America was next fastest, 15.5 days. Hawaii was the longest with almost 26 days. Of the more conventional GRs, Europe generally averaged 21 days, and West Pacific averaged 18 days. Half of the mail to the latter two GUs arrived in 18 and 15 days, respectively.

There were also some interesting F-test outcomes on these data. First, F was significant and the eta-squared value (.0369), while moderate in real terms, was relatively high within the context of the present study. There were two significant post hoc contrasts. Europe was significantly different from West Pacific. The second significant post hoc contrast had Hawaii significantly different from practically every place else. Although the evidence is not compelling, there was at least evidence that Hawaii did experience a systematic delay in course package delivery time.

Gateway Post Office. Gateway Post Office (GPO) paralleled the geographic regions discussed above, except that the distinctions had real meaning and correspond to constructs established by the MPSA and the USPS. Among the gateways for overseas mail, the fastest transit was for mail exiting via Miami (averaging 15.5 days). The slowest MPSA gateway was New York, averaging 21.4 days. Transit time for mail traveling through the USPS channel was 23.4 days. This was an expected outcome since the MPSA mail examined here traveled 4th class for only the CONUS portion, converting to the equivalent of 1st class at the gateway. USPS mail went 4th class throughout.

The differences due to gateway were apparently systematic. The $F(4,2549)$ was significant, with a moderate eta-squared (about 3% of the variability of mail transit time was accounted for by GPO). Of particular interest were the post hoc contrasts. New York was significantly different from Miami and San Francisco, and USPS was significantly different from Miami and San Francisco (New York and USPS were not significantly contrasted). A question arose why Seattle was not contrasted, even though its average transit time was faster than San Francisco. The outcome was apparently due to the fact that analysis of variance (F) does not act on the means but on the variances of the variables. In the case here, the structure of the Seattle data worked against rejection.

Type of Written Comments. By analyzing mail transit times broken down by type of written comments -- critical or non-critical -- there was an opportunity to ascertain the validity of student mail time complaints. All earlier research suggested that there was no difference in transit times between those who complained about it and those who did not. The findings of the present study were reasonably consistent with those earlier.

Only 71 of 2554 respondents (about 2.8%) actually complained about mail delay. Transit times for the complainers was about 4.5 days longer than those who had no comments to make. Curiously, though, they were only about 2.5 days longer than those with favorable or neutral comments. Median times were in the same direction.

Analysis of the variance due to type of comments was significant ($F(2,2551)=9.5601$, $p=.0001$), but the eta-squared was very low, less than one percent (.0074). There was one significant post hoc contrast, but the outcome only led to a more enigmatic situation. It occurred that "no written comments" contrasted with both complaints and positive/neutral comments, and that complaints did not contrast significantly with non-criticisms. The meaning of these data, other than the lack of a clearly demonstrated

"criticism" effect, is not clear. In the 1984 study, the complaints were clearly associated with a real, if small, increase in transit time.

Package Condition. It was known that some packages arrived damaged. Indeed, some packages did not arrive at all and the student received only the mailing label (that is, the outer envelope separated from the package). The problem of damage was (and remains) important for two primary reasons. First, it may degrade student performance. Second, damaging may delay receipt due to repackaging by postal handlers.

Almost 97 percent of all courses arrived in warehouse condition. Only about three percent were broken open, and less than half of these actually incurred damage. It was interesting to note that broken open/undamaged packages arrived faster than for other groups, averaging 19.5 days. Damaged packages took a bit longer than normal, almost 23 days. Three students received only the mailing label; in 1984 there were none in this group. (None of these the outcome differences were significantly different, however.) This was an improvement over 1984, in which the opened with damaged courses did take significantly longer to arrive than courses in the other groups.

Remarks by Type of Course. Using a contingency table, a test was made to determine whether there were significant differences in rates of comments as a function of type of course in which the student was enrolled. While the effect was small, the difference was significant. Recognizing that complaints were infrequent phenomena, it was observed that both Specialized Course and CDC students were less likely to complain than PME students (1.6%, 1.2%, and 5.2%, respectively). Equally curious, Specialized and CDC students were relatively less likely to make positive or neutral comments than PME students (13.6, 15.0, and 16.0%, respectively).

Long Transit Times and Remarks by Location. For the purposes of this study, long transit times were considered to be 35 or more days. There were 232 returns in the long time region. Of these, 168 had no comments, 51 had neutral or positive remarks, and only 13 actually complained about mail delay. As a practical matter, these were considered the only valid mail time complaints in the present survey. That is, valid mail delay complaints were received on only about engulf of one percent of ECI courses.

Detailed Analyses of Alaska, Hawaii and Puerto Rico. An intentional limitation of the 1984 study was the omission of Alaska, Hawaii and Puerto Rico as overseas locations (although they were addressed elsewhere). In an effort to broaden the comparison base, separate detailed analyses of these locations including descriptive statistics and full analysis of variance information were obtained.

Alaska. ECI courses sent to Alaska went to 14 different APOs and ZIP Codes. Most went to Elmendorf AFB in an average of 19.6 days. The second most mail went to Eielson AFB in 20.6 days. Mail times for ZIP Code 99502, an address in Fairbanks, and Shemya AFB, an APO, were interesting: 25.5 days and 15.9 days, respectively. Fairbanks is a modern urban area and Shemya is considered acutely remote; yet, Shemya mail averaged 10 days faster. None of the differences among these data were statistically significant although the eta-squared was about eight percent, suggesting that some effect may be present.

Hawaii. Hawaii was of particular interest as earlier data (not published) indicated that military official (on-base) mail was slightly faster than mail to public (off-base) locations. This opinion was not supported by the new data. ZIP Code 96853, Hickam AFB, reported an average transit time of 26 days, no different from the mean of all locations in Hawaii. Wheeler AFB, ZIP 96854, averaged 24.8 days. A military housing area (considered public in this study),

ZIP 96818, averaged 21.9 days. Other locations in Hawaii varied widely.

Apparently, there were significant differences in delivery rates in Hawaii. The $F(18,125)=2.7057$ had a probability of .0006 and eta-squared for these data was 28 percent -- very large. Unfortunately, post hoc contrasts failed to reveal any significant comparisons and the situation with regard to Hawaii became somewhat more obscure than before.

Puerto Rico. Since the Air Force had no major presence in Puerto Rico at the time, the low response rate of 24 returns was expected. These were fairly randomly distributed among seven locations. Average mail times ranged from 17 days at several locations to 35.7 days for an address in San Juan. Ramey Observatory averaged 23.5 days. Overall, the average mail time was 24.6 days. Despite the fact that a number of Puerto Rico mail times were comparatively excessive, there were no student complaints.

The F-test for these data did not reveal any significant differences although eta-squared was very large, .4782. Failure to reject was due no doubt to the low overall sample size for Puerto Rico relative to the number of locations.

Conclusions

1. Fifty percent of students received their courses in 17 days, 80 percent in 25 days, 90 percent in 33 days, and 95 percent in 43 days. These figures are uncorrected and may include time during which students were on leave, TDY, or otherwise unable to receive mail.

2. Ninety-eight percent of mail was received in fully satisfactory condition.

3. Less than three percent of the students responding complained about mail time delays. However, considerable evidence suggested that most of the complaints were not due to either the USPS or MPSA mail systems, or to any other factors under ECI control.

4. PME students were roughly three times more likely to complain about mail delays than either Specialized Course or CDC students. They were also about three times more likely to make fatal errors on the survey post card.

5. Mail time differences among locations, gateway post offices, countries, and so on, were either not demonstrated or minor. The only systematic differences involved mail to Hawaii and Puerto Rico. In both instances, however, the differences were minor and did not result in a meaningful increase in complaints.

6. These data closely conform to the findings of the last major mail transit time study conducted in 1984.

REFERENCES

- Diehl, G. E. How Long Does It Take ECI Courses to Reach Students? Gunter AFS AL: ECI Evaluation and Research Branch, Aug 1984.
- Kelly, M. I. DMS Mail Survey. Gunter AFS AL: ECI Evaluation and Research Branch, Jul 1976.
- York, J. W. Course Package Survey: A Measure of Time/Condition. Gunter AFS AL: ECI Evaluation and Research Branch, May 1980.
- and Diehl, G. E. APC Course Package Survey: A Measure of Time and Condition. Gunter AFS AL: ECI Evaluation and Research Branch, Dec 81.

INTERACTIVE VIDEO: R & D AND A PRACTICAL APPLICATION IN THE BRITISH ARMY

*by Colonel Donald H Oxley, Commanding Officer
(British) Army School of Training Support (ASTS)
and Major Dennis Quilter, Systems Consultant
(Training Development), ASTS*

INTRODUCTION

1. The (British) Army School of Training Support acts as focal point for the application of the British Army's systems approach to training. It sustains the official doctrine, by authoring a series of official publications, and by training the trainers on specialist courses. It solves training problems by the selective deployment of post-graduate-trained consultants. Its establishment includes two studios, for graphics and for video, producing work of professional quality for the Army. Finally, it undertakes R & D projects for the Director of Army Training.

2. This combination of tasks has made ASTS the natural home for the Army's exploration of the possibilities and feasibility of interactive video. Our contact with the medium goes back to 1984 when we conducted a review of the then state-of-the-art and applications. As a result of the report we published, we were tasked to investigate the potential of low-cost, tape-based systems. This phase indicated that the tape-based equipment imposed severe limitations, although simple linear programs could be constructed without too much difficulty.

3. In 1986 we were instructed to conduct a further phase of the project, using what at the time of writing the specification was state-of-the-art technology: Laserdisc, AT compatible computer, PLUTO digitized graphics system, and Kurzweil (KVS-3000) voice input device. The software provided was the enhanced TenCore system.

THE SECOND-PHASE R & D PROJECT

Aims

4. In this second-phase project, ASTS were tasked to:
 - a. Identify and assess the problems in the total process of production of IV courseware.
 - b. Assess the training effectiveness of IV courseware.
 - c. Recommend a course of action for the Army on possible training applications of IV.

It is emphasised that ASTS was required to do all the design and the software and video creation; no funds were provided for commissioning commercial assistance. The production of courseware to teach one of the subjects from an ASTS training course, Instructional Analysis, was selected as vehicle for the project.

Conduct of the Project

5. The project plan initially agreed aimed to bring together the procurement process for the equipment and the training and phased availability of personnel. The project officer was to be full-time, but the other members of the team, the subject matter expert, the training designer, the programmer and the specialist video expert had other commitments and their availability at key periods had to be pre-planned. This carefully-made plan first ran into difficulties when there were delays in procurement and difficulties over the compatibility of the various components of the system. Later, the extended life of the project meant that the original team personnel were changed because of retirements, postings and subsequent commitments of higher priority.

6. The design methodology was based upon the Structured Design Method (SDM): events which would affect the trainee progressing through the course were represented, as were the control elements for progress monitoring. In addition, a flow chart was maintained of the overall program structure and trainee interactions with the course; screen layout sheets specified in detail the information presented and control functions available. This methodology enabled the project to continue in spite of the difficulties encountered, but deadlines had to be extended and the project over-ran its projected time-scale. Ultimately, the courseware was completed to the required design, although the incorporation of the voice-interface was discontinued and made the subject of a separate report.

Costs of the Project

7. The costs of the project which produced about one hour's courseware were:

- | | |
|----------------------------------|--------|
| a. Equipment including software. | £27.5K |
| b. Video studio facilities. | £10K |
| c. Video editing and disc. | £4.5K |
| d. Project team manpower | £96K |

Total £138K

8. In reporting to the Director of Army Training we concluded that, because of the high setting-up costs and the need to allocate a dedicated team to production over a long period, IV is, unlike CBT, not yet amenable to in-house production by training units of the British Army. More positively, we recommended that IV should continue to be considered as a possible (although exceptional) solution to training problems, and that the particular areas of application which appeared suitable were simulation and inter-personal skills.

9. It was soon after submitting these recommendations that we found ourselves asked to solve a training problem, and in due course recommending an IV solution.

THE THERMAL IMAGING RECOGNITION PROBLEM

10. A number of different Thermal Imaging (TI) sights have been or are being introduced into the British Army. These devices are employed in surveillance, acquisition of target, and target engagement. In each of these roles there is a requirement to identify what is observed as a basis for subsequent tactical decisions. Crucially the recognition of targets as 'enemy' precedes weapons engagement.

11. Armoured Fighting Vehicle (AFV) and aircraft recognition training in the normal visual mode is currently conducted in the Army. However, because thermal images are very different from visual images, particularly at medium and long operational ranges, there is no natural transfer of recognition ability from normal to TI images.

12. This problem was highlighted by the user who gave evidence of a skill/knowledge deficiency by declaring he did not have the ability to identify potential TI targets on the battlefield. Consequently the Director of Army Training tasked ASTS to study the problem and make recommendations as to the best methods of training.

13. The study commenced with no pre-conceptions with regard to solutions and employed a traditional needs analysis approach. Some training in TI recognition was being conducted at this stage as Army training establishments responded to the problem. This training was, however, varied across different users and at an embryonic stage. Industry was also aware of a potential market for TI recognition training product support materials (and therefore profit) but was hampered by the dearth of source material which mainly lay in the hands of the military. Clearly a methodology had to be applied which would permit the evaluation of existing and proposed solutions and also give a spur to future development. It followed that the training requirement should be derived via Front End Analysis (FEA) techniques as an adjunct (in some cases rather belated) to the main equipment development cycle. The overall methodology adopted for the study is shown diagrammatically at Figure 1.

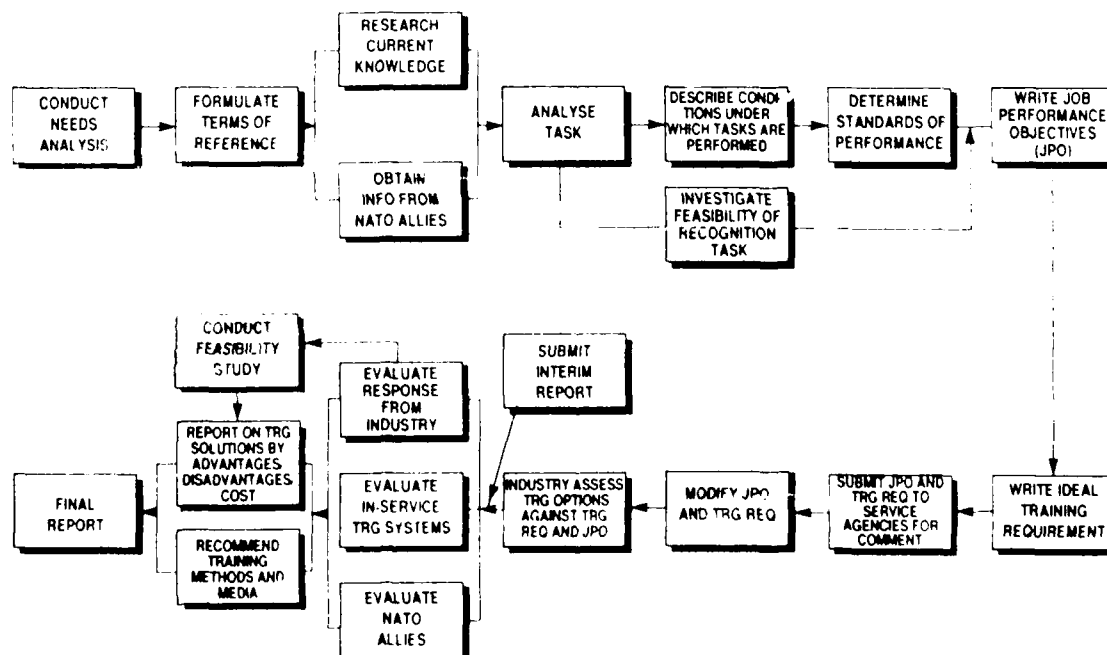


Figure 1. ASTS TI RECOGNITION TRAINING STUDY-METHODOLOGY

14. IV became apparent as the most viable form of training once the required learning design elements had been derived from the following:

- a. Thorough task analysis and definition of conditions and standards under which the job was performed.
- b. An investigation into perceptual recognition theory.

15. From the above analyses the desired learning design elements were:

- a. Maximum trainee exposure to images to be learned thus providing the opportunity for multiple associations and discriminations to be mentally internalised. (Reed 1972, 1978) (Ferry 1986)
- b. Moving imagery both as reflecting task analysis and to provide an almost infinite number of views on image rotation (a static view represents only one 'frozen' image and is thus not learning efficient). (Wickens 1984, Merrill and Tennyson 1977)
- c. Realistic imagery in terms of the full spectrum of conditions envisaged on the battlefield (eg range, aspect, meteorological conditions). This determines that the trainee must have the facility to manipulate the contrast, brightness and polarity of images as defined by task analysis. (NATO TCP 1981)
- d. Variable ability of operators in the recognition task suggested that training should be against both a minimum standard and a potentially higher 'personal achievement standard'.

16. These learning design elements were then assessed against possible training options (eg slides, CBT, models) in a training effectiveness analysis. The best match was found to be IV where the desired learning elements are closely associated with the following delivery characteristics of the technology:

- a. Rapidly accesses/displays/ realistic images from a large library held on disc.
- b. Allows the trainee to learn at his own pace.
- c. Provides opportunities for the trainee to review the imagery either of his own volition or via system branching and remedial loops.
- d. Provides dynamic formative and summative assessment.
- e. Provides a record of performance both on floppy disc and print out.
- f. Motivates the trainee via immediate feedback on performance.
- g. Can exploit variable ability and motivate the trainee by associating a points score with trainee performance; a minimum score being set as a pass standard.

From the organisational viewpoint IV was seen to be attractive as it was capable of delivering a common standard of instruction to widely dispersed units for both initial training and maintenance of training. It could also provide print out analyses of operator performances for both instructors

and employing officers to be set against Army Training Directives. Additionally, the data base, providing both videoed TI and synthetic imagery, could be modified and updated from a central image generating facility, pressed onto disc and distributed to units.

17. Independent confirmation of IV as a solution came from industry who had been asked to recommend training methods based on a draft requirement.

18. As is emphasised in the full project report, initial development was of a part task trainer to train operators of TI devices in classification, recognition and identification skills. Other elements identified in the task analysis (eg battle picture cues, operator change of field of view) were recommended for subsequent incorporation in higher mission training

THE IV FEASIBILITY STUDY

19. Before moving to full procurement a feasibility study was undertaken to test the theoretical case for IV-based recognition training. This project was funded under the acronym TIRT (Thermal Identification and Recognition Trainer) and implemented by a team consisting of:

- a. A Service agency responsible for supplying 'rough' edited TI imagery.
- b. Subject matter experts (from the Infantry, Royal Armoured Corps, Royal Artillery and Army Air Corps) responsible for validation of imagery and input to instructional design.
- c. A specialist IV firm (SAV Communications Plc) responsible for fine edit and preparation of data base, and software design.
- d. A training consultant (one of the co-authors) responsible for instructional design, formulation of TOs and the coordination of all aspects of the project.

The user involvement in data base preparation and instructional design ensured compliance with the operational task and original training need and the use of commercial services was designed to overcome the problems encountered in the Second Phase R & D Project.

20. The evaluation has been carried out in three phases. Phase one employed a small representative sample and allowed minor bugs to be removed from the program. Phase two took advantage of a large opportunity sample by offering TIRT as a background activity to over 900 all ranks attending the 1988 United Kingdom Land Forces MILAN Anti-tank Guided Weapon Concentration. Objective data was taken from the program assessment of trainee performance and subjective data concentrating on motivational, satisfaction and value factors was obtained by questionnaire.

21. Phase two tended to confirm TIRT as a viable training system but it became apparent certain aspects of the design specification had not been met. Additionally it was found difficult to control the evaluation in an exercise environment that had field deployment and live firing as its priorities.

22. Following work by the contractor to bring TIRT up to specification the Phase three evaluation was set up. This employed a tighter experimental design as follows.

Gp A 15 trainees	X_1	X_2	0	X_2	X_1
Gp 2 15 trainees	X_1	-	0	X_2	X_1

Where X_1 was a test of normal visual recognition ability using the same AFVs as in the TIRT final assesment.

X_2 was the TIRT final assessment
0 was the TIRT learning treatment

EVALUATION FINDINGS

23. The concept of IV recognition training is strongly supported by the subjective data, based on a representative sample of 112 trainees. For instance, 94% of trainees regarded the TIRT training experience as either 'very valuable' or 'valuable'. Seventy three per cent believed that the issue of a fully developed TIRT would lead to a 'greatly improved standard' in units and 22% believed that an 'improved standard' would result. All of the trainees considered TIRT to be a better form of training than instructor led training justifying this view mainly by reference to the self-paced nature of the program.

24. **Objective data** Because of the problems with the program and experimental control described above and because the phase 3 evaluation is not yet completed it is only possible to report indications from the objective data in this paper. The average improvement in identification ability currently recorded is 50% although the standards laid down in the training objective, as derived from the subject matter expert, have not always been met. However, the standards laid down must be regarded, to some extent, as arbitrary and trainees were often not competent in normal visual recognition ability before commencing TIRT training; a design criterion of the program. Additionally, the evaluation has illustrated deficiencies in the trainer, whose incorporation are likely to improve trainee learning efficiency. For example, detailed analysis of trainee performance highlights confusable sets and permits a remedial strategy to be tailored for each trainee. Currently hard copy raw data is provided via the program, but the analysis can only be done by an instructor and no data exists at present to show the effect of such remedial strategies: say via re-teach and repeat assessments. The design specification for full development TIRT will incorporate this aspect in the student management system together with other lessons learned.

25. **Subjective data.** Space does not permit an exhaustive report of the evaluation findings to date but, in summary, the subjective data provides strong support for IV recognition training and current objective data tends to confirm the validity of the concept. However, final conclusions must await the completion of the phase 3 evaluation.

REFERENCES:

- Ferry, G. Parallel Learning in Brains and Machines. New Scientist 13, March 1966.
- Merrill, D. and Terayson, R. Teaching Concepts. An industrial Design. Ed Tech. Publ. Englewood Cliffs N.Y. 1977.
- Reed, S.K. Schemes and Theories of Pattern Recognition in Handbook of Perception I (1970). Academic Press.
- The Technical Cooperation Programme (TCP) Perceptual Recognition Training in the Military Context. Report by Technical Panel UTP2 (Training Technology) November 1981.
- Wickens, C.D. Engineering Psychology and Human Performance. 1984. C.E. Merrill Publishing Company, USA.

SIMULATION, CBT & TESTING -
COMPROMISE, CONFOUND OR CAMP?

by

Cdr. Robert H. Kerr, CF
Maj. Marilyn Hoggard, CF
Capt. Bruce D. Hyland, CF

Canadian Forces Fleet School Halifax

Introduction:

The relative discrepancies between training and operational mastery levels determined by the systematic analysis of job requirements often show themselves in the differences of opinion regarding the degree of fidelity required to reach desired operational performances. In the same sense, the designer of a predictive measurement instrument faces much the same problem, i.e., what will predicate the device's design - a measurement instrument that parodies the operational performance or success of candidates, or alternatively, a measure of the traits or sub-traits identified as a function of the operational analysis? The authors have ambitiously compared the relationships involved in the psychometric and training arenas - this comparison has clarified some of the issues of importance in overcoming deficiencies in communication. The relationships between fidelity and validity have indeed proven quite useful in the analysis of the training requirement. Initially, the concept of fidelity will be briefly discussed.

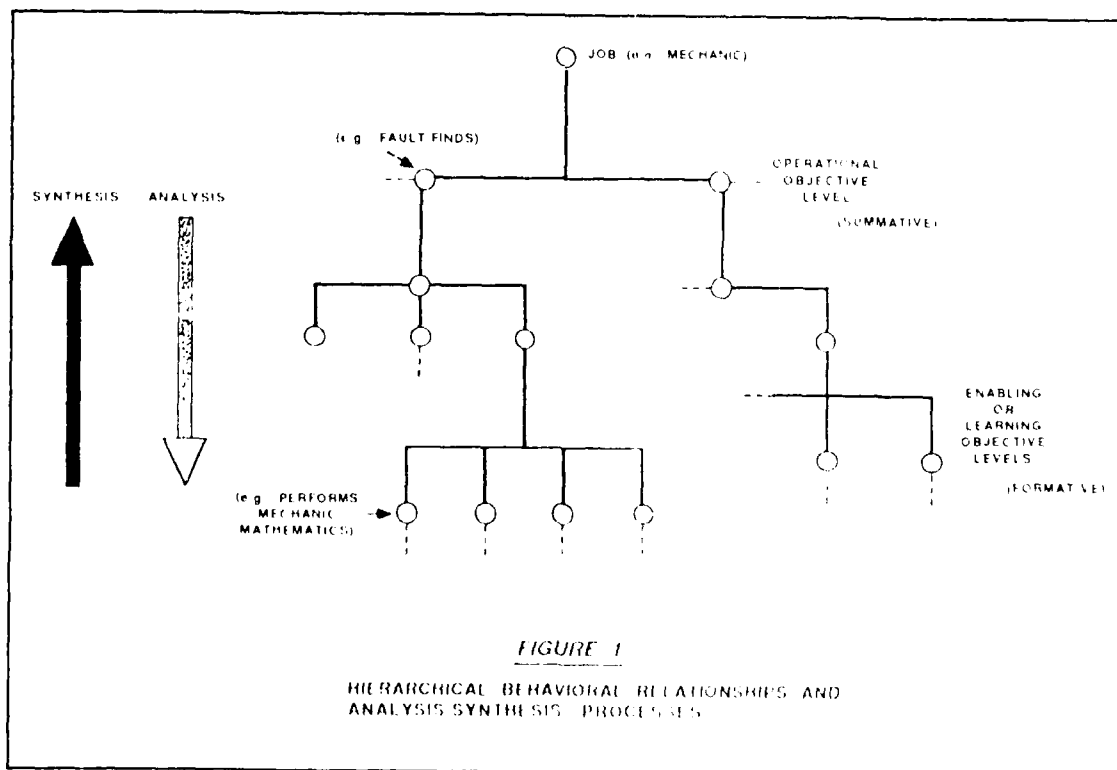
Facets of Fidelity

In its simplest form, fidelity has been defined as the 'degree of representation' a device possesses. The 'degree of representation' can be expressed in many ways, however the fundamental representative characteristics appear to be those of functionality and physical similarity to the operational environment, Bruce (1987). From a measurement point of view the degree of similarity or fidelity is usually indicated by a percentage or index of replication. The authors have used a simple percentage ratio of the number of skill related tasks possible on a device or attainable in a training programme divided by the number of operational tasks the graduate will be called upon to perform. Traditionally, the argument regarding 'how much' fidelity is required in a device centers around the cost-effective payoff of the device vis à vis using operational aircraft or ship systems, Behavioural Team et al, (1988). Because the costs of replicating operational environments are plummeting, there is some danger that such technological advancements will be used inappropriately in a training programme, without due regard to the learning process. In other words, failure to recognize the requirements of the learning process by striving to acquire full-fidelity systems without due attention to learning stages (or enabling stages) may result in a less than satisfactory training programme. (For an extreme manifestation of this see Frank et al (1983).) It is the authors' contention that different stages of a training programme themselves demand different 'types' of fidelity.

Longitudinal Aspects

In psychometric usage, the equivalent of fidelity appears to be quite simply, face validity. If a parallel could be drawn, many predictive tests do not necessarily exhibit this type of validity at all, having been derived for reasons of useability (Popham, 1975, P. 122) measured indirectly by means of concurrent validity indices (a comparison of the Miller Analogies test with SAT is a good example). In the same vein, the training designer or analyst is concerned with developing a training programme which is efficient in the learning process - the shorter the better - as long as the prescribed skill levels are achieved in the terminal stages.

What emerges from these comparisons is simply that in a student's or test's formative periods considerable emphasis is placed on efficiency - can the test predict efficiently or can the student learn efficiently? The authors have categorized the requirement for the training situation to show 'formative fidelity' which can be defined as the degree by which the training programme's enabling stages (strategies and enabling objectives) represent efficient methodologies of learning. This type of fidelity can be measured by the experimental comparison of a training programme with alternative training methodologies and by comparing the similarity of actual achievement test devices at these stages with the stated enabling requirements. In predictive tests, the equivalent type of validity might be the construct validity of tests to indicate a domain's traits or sub-traits that are attempted to be measured. At the formative stages, therefore, it can be inferred that full job, or face validity to the composite criterion may be unnecessary - the requirement at this stage being one of efficiency.



The latter stages of a training program on the other hand present different requirements. Recall that the training analyst, through an orderly 'breaking down' of the composite performances has derived enabling and learning requirements to build the training strategy. When a student reaches the final stages of the programme, a requirement to synthesize these earlier activities leads to a 'summative' conclusion where the traditional meaning of fidelity comes into play. At this stage, full and accurate representation of the operational situation is desirable. In the authors' opinion, therefore, the requirement here is for the highest degree of 'summative fidelity'. Figure 1 displays the above relationships in a training programme structure. In predictive testing, the actual test forms used may or may not show face validity or summative fidelity, however to be effective the test must show high criterion-related (primarily predictive) validity as well as convergent validity with the criterion, Sax (1980).

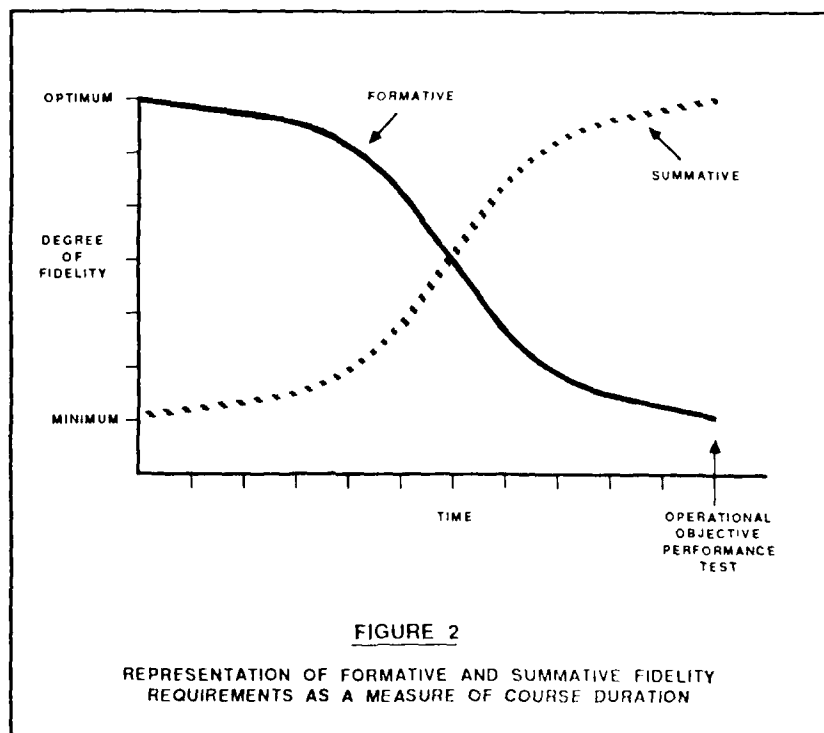
Implications

Making comparisons is as fraught with danger as is the provision of examples or analogies. The authors have used the categorizations of summative and formative fidelity as a framework for discussion when dealing with simulator manufacturers, training analysts/designers and evaluation specialists. Because the stated requirements for each stage in the learning process are different, different measures are taken to assess the effectiveness and efficiency of each stage.

Computer-based training (CBT) programmes, for example, provide the analyst with a rich field to apply these fidelity criteria. The instances in the recent literature specifically regarding CBT testing standards, item construction related to type, contamination issues, e.g., Sarvella and Noonan (1988), indicate that considerable work and emphasis is being placed to increase the formative fidelity of CBT, enabling more efficient learning to take place. Although not a direct comparison of two methodologies, researchers Kloosterman, Harty and Matkin (1988) have studied relative computer based useage in a mathematics curriculum in categories of simulation (includes practice) and learning (new materials, develop problem-solving or "higher level" thinking skills). In reporting that "Simulation and Problem-Solving software forces students into critical thinking about topics not frequently addressed before computers were available", our assessment might be that the developing summative and formative fidelity of such systems are improving, enabling the attainment of behavioural objectives which are difficult to achieve through the means of "lectures or the use of printed materials." (P.46)

The relationships between formative and summative fidelity open up some fascinating avenues for future investigation. In training contexts, where the behavioural objectives can be quite distinct, and the learning methodologies more easily evaluated and hierarchically structured, the requirements for the fidelities discussed might take the time profile of Figure 2, where the requirements for each fidelity changes over time in the programme. Requirements for summative fidelity at the start of the programme are infrequent (motivational, attitudinal, situational etc.) and grow throughout the programme until the requirements for integration, consolidation, drill and practice take precedence. On the other hand formative fidelity requirements initially are high in frequency and gradually diminish as the programme reaches incrementally small 'learning progress steps' near the end of the programme. (Figure 2 has

been provided for discussion purposes only and does not represent any experimental or empirical results.) As a passing comment, the effective use of job-aids may completely avoid the need for the provision of formative fidelity at all.



In educational environments on the other hand, where objectives are essentially divergent, unrelated in great measure and less hierarchically homogeneous, the relationship between the two fidelities could take the form of many aggregate relationships of Figure 2 for all behavioral objectives in the programme.

Finally, no discussion related to fidelity or simulation is germane without the mention of transfer of training. Fundamentally, a training programme should strive to maximize transfer by exhibiting high formative and summative fidelity. It is also clear, however, Morse et al (1988), that transfer must optimize other factors as well, which can be totally unrelated to these indices: environmental acceptance and policy for example. With this in mind the 'transfer equation' could be:

Degree of
Transfer = $f(F_F + F_S) + f(E_P)$

where $f(E_P)$ is the function of the "Environment of parameters" involved in the

F_F = degree of Formative fidelity

F_S = degree of Summative fidelity

Summary

The introduction of the terms formative and summative fidelities has provided useful contexts for communication, assessment of training methodologies, determining specific requirements of training programmes and specifying hardware, software and courseware needs. Further research should provide for more accurate measurement of both types of fidelities and lead to a better understanding of the principles involved in their interrelationship.

References

- Behavioural Team. (1988). Functional training requirements report: CPF/SRP II trainer study. (Contract No. W8479-7-CA05). Ottawa, Ontario: National Defence Headquarters.
- Bruce, R.C. (1987). Simulation fidelity: A rational process for its identification and implementation. Proceedings of 9th Interservice/Industry Training Equipment Conference, pp.20-25.
- Frank, L., Kennedy, R.S., Kellogg, R.S. & McCauley, M.E. (1983). Simulator sickness: A reaction to a transformed perceptual world: Scope of the problem. (Report No. EOTR 88-2). Orlando, FL: Naval Air Systems Command.
- Kloosterman, P., Harty, H. & Matkin, J. (1988, October). Computer utilization in elementary school mathematics classrooms. Educational Technology, pp.42-47.
- Morse, S.L. & Nudell, D. (1988). Transfer: A critical factor in training. Project Report: SJNS 8801, (Contract No. 8806). Halifax, Nova Scotia: Canadian Forces Fleet School Halifax.
- Noonan, J.V. & Sarvella, P.D. (1988). Implementation decisions in computer-based testing programs. Performance and Instruction Journal, 6, pp.5-13.
- Noonan, J.V. & Sarvella, P.D. (1988, May). Testing and computer-based instruction: Psychometric considerations. Educational Technology, pp.17-20.
- Popham, W.J. (1975). Educational Evaluation. Englewood Cliffs, NJ: Prentice-Hall.
- Sax, G. (1980). Principles of educational and psychological measurement and evaluation, (2nd ed.). Belmont, CA: Wadsworth Publishing.

Applications of Simulation and Wargaming To Training¹

Franklin L. Moses
U.S. Army Research Institute
for the
Behavioral and Social Sciences

and

Earl A. Alluisi
Office of the Secretary of Defense
(Research and Advanced Technology)

Simulation-based wargaming and training, properly employed, can provide practical and affordable means of raising the levels of readiness of our military forces to new highs. The use of computers for developing and controlling training scenarios has, in recent years, provided increasing training capabilities. More importantly, research and development (R&D) on applications of computer technologies to training provides increasing promise of substantial improvements and extensions of cost-effective combat-mission training capabilities, both vertically (i.e., from theater-level commanders and chiefs--or CINCs--down through the various command and staff organizations, to the individual crews of combat vehicles such as tanks) and horizontally (i.e., through the different staff offices at a given level).

The Defense Science Board in 1988 approved recommendations to the Chairman of the Joint Chiefs of Staff, made by its Task Force on Computer Applications to Training and Wargaming, to provide simulation-based wargaming and training capabilities for the CINCs, their staffs, and their component commanders, staffs and units. Such training, it was pointed out, would benefit U.S. joint forces by providing realistic opportunities to exercise the existing command-and-control procedures, to develop improved procedures, and even to examine "what if" war plans, in coordinated training with commanders and units that would be assigned in time of war.

"Joint" in the present context means simultaneous, operational employment involving different U.S. forces from two or more of the Services. The Task Force's report provided a broad plan on how best to apply advanced computer technologies and improve joint force training and doctrine. One of the major findings was that there is currently little or no joint training. Most of the training is single-Service oriented, and even where a second, third, or fourth Service is represented in a wargaming model, the representation is for purposes of impacting the single-Service training, not for jointly training the Services. The present paper focuses on the potential use of wargaming for training at the higher echelons of command, and on the emerging roles of computer-based simulations in supporting such wargaming and training.

¹The second author presented and published a related paper entitled "Wargaming: Applications of Human Performance Models to System Design and Military Training" as part of "Applications of Human Performance Models to System Design: A Technology-Demonstration Workshop," NATO Defense Research Group (Panel 8), Research Study Group 9 Workshop, Orlando, FL, 9-13 May 1988.

WHAT IS WARGAMING?

A wargame is a simulation of a military operation, involving opposing forces, using information designed to depict actual or real-life situations.

Some wargames are fully automated; i.e., they are computer simulations that run without human intervention. Other wargames are interactive; they involve man-in-the-loop simulations. Interactive wargames differ in the degree of human intervention permitted or required.

Wargaming simulations have been constructed, demonstrated, and used to represent a broad range of situations. Some represent very large global and theater-level conflicts. Others represent smaller sectors of larger conflicts, such as a naval battle group or an army corps air-land battle zone. Still others, designed especially for training, represent smaller collective battle elements such as an army tank battalion, company, or platoon, or even individual battle elements such as a single aircraft, ship, or army tank.

There is increasing interest in the networking of wargaming simulations to provide for the interactive, and potentially quite realistic, training of collectives (crews, groups, teams, and units, or CGTUs) as well as of individuals.

USES OF WARGAMING:

There are two primary purposes of interactive wargames: (1) military education and training, and (2) battle-management training and aiding. Military education and training is meant to provide a broadening experience that replaces ignorance with knowledge of doctrine and procedures. That is the province of the senior Service schools and colleges. The purpose of training, however, is to provide for acquisition, refinement, and maintenance of battle-winning skills. Battle-management training, in all services, is regarded as the province of the operational commander. Educational experiences may be infrequent in a lifetime, but training experiences have to be sufficiently frequent to support not only the initial acquisition, but also the further refinement and maintenance of the skill at a high level. Until recent years, wargaming has been used relatively effectively for education, but nearly not at all for training. The high cost and limited availability of wargames for training have been factors that technological advances are beginning to overcome.

As technology advances, particularly in computer and communication technology, the uses of wargaming in battle-management training and aiding is increasing. The prediction is that this will be the fastest growth area in future wargaming development, with an increasingly "fuzzy" distinction between "aiding" and "training," and a growing realization that this technology provides a most cost-effective contribution to "readiness". Wargaming should be regarded as providing a real force-multiplier effect.

POTENTIALS OF WARGAMING IN TRAINING:

Is there evidence that the provision of wargaming simulations for training would produce beneficial results? There is evidence in the open literature. Some of it is direct, but much is indirect. For example, during 1987, U.S. Army tank platoons stationed in Europe were trained with a network of tank simulators. The skills they gained in the simulator training transferred

positively to the NATO-exercise range, and the U.S. Army tankers took first prize in the annual competition. This was their first win in ten years.

That there are vast differences in the battle-effectiveness of commanders and their units is a widely recognized fact among both warriors and military historians. Naval commanders and historians know of "good squadrons" and "bad squadrons," identically equipped and often fighting in the same area. Data are not impossible to find. For example, one NATO-sponsored study of WW-II German U-boat commanders found that 33% failed to engage any targets, and 13% failed to hit any targets they found. Thus, 46% were ineffective. On the other hand, the best 10% of the commanders sank 45% of the allied ships sunk. Such differences are usually explained with reference to some combination of the "human factors," e.g., selection and training, both of which can be impacted by wargaming and training.

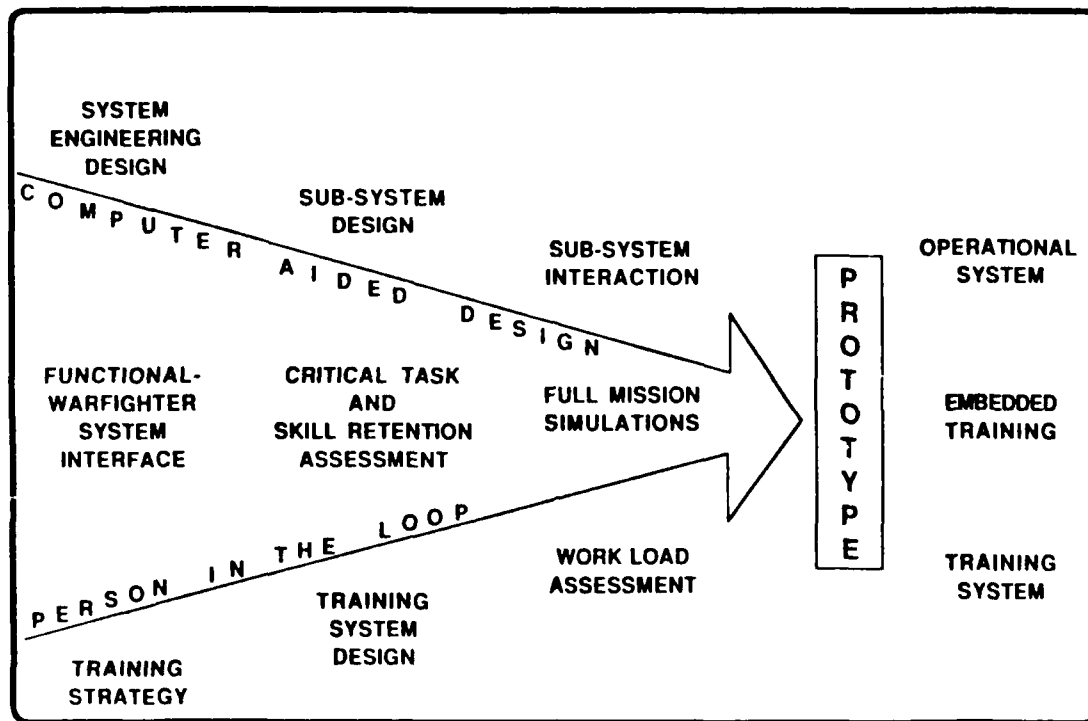
A third and final example is drawn from the WW-II air-battle experiences of the German and American forces in the European theater. In air combat loss rates, the German pilots experienced near 40% first-mission losses, whereas the American first-mission loss rate was half that, or 20%. Essentially, the Americans had a five-mission advantage relative to the Germans. This advantage has been attributed fully to the differences in training. The Germans were short of fuel, and although they trained their pilots well to fly the aircraft, they did not have fuel "to waste" on the training of air-combat maneuvering skills. The Americans did and the results are obvious. Evidence obtained with the Advanced Simulator for Pilot Training (ASPT) at the Operations Training Division of Air Force Human Resources Laboratory has more than suggested that similar advantages can be gained by augmenting flying training with air-combat simulator training.

There is sufficient evidence, in theory as well as fact, to indicate that combat experience is the best trainer of combat skills. Combat veterans have out-performed "green" troops in every recorded engagement. The technological capability is now present to be provided in training systems that can produce combat veterans and aces in peacetime! Can that capability be prudently ignored? If not, how can the technology be applied to the training of combat-relevant skills?

CHALLENGES:

The technical challenge of using wargames for training is to develop adequate degrees of realism and fidelity to provide effective "war-fighting rehearsal". Figure 1 shows a schema wherein system engineering design and training strategies using simulations are developed together so that training experiences may be as similar as needed to operational experiences. The figure emphasizes how the development of simulations must integrate the system design process shown along the top of the arrow with the human resources and training represented along the arrow's bottom. The result is a continuum ranging from an operational system, through training that is incorporated or embedded into the system, and ending with a separate training system. The ideal wargaming system would identify and develop the advances in computer and networking technologies for simulations that would provide affordable augmentations of field exercises. Such wargaming also must be frequently available since one or two exercises per year will do little to hone "fighting skills".

FIGURE 1. TRAINING APPLICATIONS SCHEMA



The technologies available for use in the development of training systems are shown in Figure 2. We tend to have most knowledge about those at the top left and less as we progress down the list to the bottom right. The challenge here is how to exploit these technologies in a way that is consistent with the training needs of commanders, staffs, and units.

FIGURE 2. WHAT TECHNOLOGIES

PROCESSORS/MEMORIES	OPERATING SYSTEMS
DISKS	INTERFACES
GRAPHICS/DISPLAYS	PROGRAM LANGUAGES
COMMUNICATIONS	EXPERT SYSTEMS
DATABASES	SIMULATIONS

Those technologies at and near the top generally are hardware-based where the risk in training applications is lowest and progress is greatest (in terms of increasing capability and reducing costs). These are truly the enabling technologies. Those technologies at and near the bottom of the list -- including simulations -- generally are the "cognitive-based" human technologies for

which the risk in training applications raises great promise but progress is questionable at best. These are truly the limiting technologies at present.

OPPORTUNITIES AND RECOMMENDATIONS:

Military "readiness" depends in great part on the battle-management (tactical and strategic) skills of the commanders, staffs, and "troops" -- both as individuals and CGTUs (i.e., crews, groups, teams, and units). Their skills are determined by the relevant experiences they have had in learning to be effective commanders, smoothly coordinated staffs, and battle-wise warfighting "troops." In time of war, it is a truism that combat-veteran commanders and CGTUs are better -- much better -- than "green troops." In peacetime, military training has the goal of providing the experiences relevant to the acquisition of such skills so that the first combat engagements are successes.

Simulation-based training is an essential component to field and sea training in providing such experiences. The technical opportunity is here. The task is to identify and develop further the advances in computer and networking technologies so that systems allow different simulations and wargames to be used to provide the necessary training opportunities -- simultaneously, interactively, and both horizontally (among commanders and staffs at the same levels) and vertically (among commanders and troops at subordinate levels).

The current opportunities are further enhanced by other circumstances: (1) the Services appear to be ready to accept and use such training, (2) the technology is here and coming, driven primarily by developments in a vast civilian market, and (3) the recent legislation that strengthened the role of the Organization of the Joint Chiefs of Staff in the matter of joint training authority and responsibility.

In this regard, the Defense Science Board Task Force recommended that the first step should be the internetting of the existing wargaming assets, for example of the senior Service schools and colleges. They recommended that with the internetting, plans be made to play in month 12 the best joint war-game that could be arranged. They proposed that by doing so, a baseline would be established from which other improvements and enhancements could proceed. They advised to work towards interoperability but cautioned that the first step should be the internetting of wargaming and simulation capabilities through communication technology.

ISSUES YET TO BE RESOLVED: R&D CHALLENGES

There are numerous issues yet to be resolved, many of which require further information that the R&D community can provide. In point of fact, experiences (including lessons learned from both successes and failures -- or, at least, partial failures) are needed in how to design, build, and use such simulation-based broadly applicable training SYSTEMS, the component parts of which include the computer hardware, the models, wargames, simulations, etc., and the network itself. The need for a tested and effective INTERACTIVE NETWORK is the greatest current challenge, for that is the key element. Networking is the enabling technology. Simulations with different characteristics, from an Army commander's view of a mechanized battle to an Air Force commander's concern with close air support to the Navy/Marine warfighting role must be capable of being "hung" onto the network, and of interacting, in the great numbers of weapons and vehicles that would be necessary for realistic combat representations.

Another major challenge is for measurement and evaluation of the performances, and the timely provision of appropriate feedback, so learning can take place and both training and the test and development of new tactical doctrine can be effective. Some progress has been made using a methodology called the Headquarters Effectiveness Assessment Tool (HEAT) by the Navy and the Army Command and Control Evaluation System ACCESS in a modified version. The technique has successfully been applied to force-level (e.g., Task Force, Brigade, Division, Corps) battle-staff evaluations to identify, quantify, and assess critical performance factors in information flow and tactical decision making in a field setting. It is especially important that more progress be made in the measurement and evaluation of the performances of CGTUs, and not merely of individual performance.

THE PROMISE AND THE GOAL:

All this comes together as a goal -- a goal that promises a peacetime training capability to provide commanders, staffs, and troops that are truly at the highest level of combat-skill readiness -- a level equivalent to that of battle-seasoned warriors. Already, a graphic-based simulator like the one in SIMNET, an Army simulation network that began with tank training stations, shows the promise of interactive networks. With appropriate distribution of simulation-based wargaming and training, we may practice wartime capabilities using realistic vehicle battle stations as well as command-and-control activities for individuals, groups, teams, and units. The research and development community has the opportunity to participate in (1) integrating hardware, software, and human technologies to improve military command and control, effectiveness, and readiness, and (2) establishing a testbed and a rapid-prototyping program to provide simulation and training techniques. The enabling technologies are ours to develop.

COMPUTER BASED TRAINING (CBT) - A NAVAL TRAINING EVALUATION

Authors: Lieutenant Commander Peter J ROSS Royal Navy
Lieutenant Commander S L LATIMER Royal Australian Navy
Lieutenant Commander A R JONES Royal Navy

The Royal Naval School of Educational and Training Technology (RNSETT)
HMS NELSON, Portsmouth, Hampshire PO1 3HH, ENGLAND

INTRODUCTION

1. As a result of the 1981 UK Defence White Paper there was a requirement to reduce the manpower and resources used in Royal Naval shore training and to transfer as much training to sea as possible. Against this background the Ministry of Defence decided in 1984 to investigate the use of CBT as a means of improving the effectiveness of Naval Training and enable it to be delivered afloat.

2. The Ministry selected 6 training problems on which to trial CBT. The aims of this trial were to identify effective procurement procedures for bespoke CBT systems and to determine the training effectiveness of the medium. This presentation is concerned with the training evaluation phase of the study.

STRATEGY AND CONDUCT OF THE EVALUATION

3. The RNSETT was tasked with determining the training effectiveness of the CBT which had been produced for the study. It was decided that the evaluation should address those issues which were likely to impact on future CBT policy and the aim of the evaluation was therefore defined as:

"To evaluate the 6 pilot CBT projects with regard to the efficiency and effectiveness potential of CBT as a means of providing training both ashore and afloat".

4. A 'top-down' strategy was devised to ensure that all the relevant issues would be addressed and the upper levels of this are shown in simplified form in diagram 1. By extending this approach the questions posed at each level became progressively less subjective and ultimately identify the required data. The following supporting objectives were defined using this strategy:

- a. To determine the instructional effectiveness of CBT.
- b. To assess the degree to which CBT equips trainees for their operational job performance.
- c. To assess the attitudinal response of trainers and trainees to CBT.
- d. To investigate the effects of CBT on the organisation of training.
- e. To investigate the impact of CBT on human and material training resources.
- f. To forecast the implications of the wider applications of CBT both ashore and afloat.

5. Performance and attitudinal data were collected between January 1987 and February 1988. Papers addressing the issues defined by the study's supporting objectives were produced for each individual project. The final report drew on these findings to make general conclusions and recommendations about the utility of CBT in the Naval Training Environment.

TRIAL LIMITATIONS

6. Manpower and funding constraints led to the following limitations on the evaluation:

- a. It had to be assumed that the CBT being evaluated was appropriate, of good quality, employed effectively, and was the best possible response to the training problems selected.
- b. Much of the courseware trialled was tutorial (63%), with some drill and practice/emulation (28%) and only a little simulation (4%). This restricted spread of modes precluded a full evaluation of the potential of CBT.
- c. Changes in trainee performance were often hard to quantify for the following reasons:
 1. The amount of material converted to CBT was small compared to the whole course.
 2. The validity of some pre CBT assessment tests was questionable, possibly due to the difficulties experienced in overcoming the stated training problems. Probably as a consequence of this some of the available pre CBT test scores showed little headroom for improvement.
 3. Since the evaluation was instigated after CBT entered service it was often difficult to obtain sufficiently detailed records of trainee achievement prior to the introduction of CBT.

Notwithstanding these limitations the evaluation provided much valuable information and has been instrumental in shaping RN CBT policy.

FINDINGS

7. The evaluation was designed to draw global conclusions about the utility of CBT in the Naval Training Environment and the findings are outlined below:

- a. Instructional Effectiveness. Tutorial CBT produced no significant increase in trainee performance as measured by test scores. In the more successful tutorial applications there was, however, a widespread feeling that trainees had gained a deeper understanding of the material, that their retention was improved and that they were therefore better motivated. There was also some evidence to suggest that greater consistency of performance had been achieved. In the less appropriate applications, benefits were possible but only where there was sustained commitment and increased effort by the training staff. In the case of drill and practice/emulation applications, however, there were significant and quantifiable increases in trainee achievement.
- b. Job Performance. Due to the time limitations imposed by the trial it has not been possible to effectively assess the impact of CBT on job performance.
- c. Attitudes. Tutorial CBT was well received only where the courseware was of a high standard and, even then, trainees were wary of being over exposed to it. Instructors appreciated the consistency offered by the medium and the fact that they were freed to concentrate on the weaker trainees. Where tutorial CBT totally replaced traditional instruction, it was found to be unacceptable and its use has been discontinued. Drill and Practice CBT has been enthusiastically received by training managers, instructors and trainees.

d. Organisation. The effects of tutorial CBT on training planning were either insignificant or adverse. Where classes are programmed to a fixed timetable it has not been possible to capitalise on any savings made from the self-pacing facility of CBT. In drill and practice applications, increases in trainees achievement can be directly attributed to CBT's self-pacing capability and savings in training time may be possible.

e. Training Resources. Tutorial CBT, if carefully selected and appropriately applied, offers an alternative instructional strategy and increased consistency of instruction. It has not replaced any instructors but, in the more appropriate applications, it has allowed them to be employed more effectively. As the tutorial CBT was largely used to supplement, rather than replace the traditional training materials, there were no cost savings to offset the initial costs of CBT. In drill and practice CBT, self-paced learning proved a highly successful strategy and there were significant resource savings in the following areas:

Backclassing - this was reduced by up to 80%.

Automation of manpower intensive marking tasks.

Instructional preparation time - this was virtually eliminated.

Teaching space - this was significantly reduced due to the compactness of the CBT hardware compared to the traditional teaching aids it replaced.

Management data - comprehensive statistics on trainee achievement and rate of progress were readily available at virtually no cost. This allowed early identification of trainees experiencing difficulties and the flexibility of CBT provided effective remedial training.

A cost benefit analysis of the drill and practice Morse Code training CBT showed that, on the previous year's trainee throughput, the additional capital cost of the CBT equipment compared to a traditional replacement solution would be recouped after 17 months.

f. Wider Applications. Indications are that tutorial CBT would not be effective at sea for stand alone basic training. The transportability of the material is likely to be difficult if wider applications for courseware are not considered at the time functional specifications are produced. Drill and Practice CBT is generally seen as having wider applications, especially for continuation and refresher training.

The evaluation also assessed existing applications of CBT in Royal Naval training, and it was found that these experiences generally supported the study findings.

GENERAL FINDINGS

8. Production of good bespoke commercial courseware is expensive and very time consuming. It is clear that the functional specifications need to be fully detailed and based on a thorough training needs analysis. The success of the courseware is directly proportional to this effort and its quality heavily dependent on Service input of subject matter and training strategy expertise. In one project, approximately 48 man weeks of Service effort at officer/senior rate level was invested in the production of 6 hours of courseware; where such investments were not possible this was reflected in the quality of the material. Wider applications of any courseware should also be considered at the functional specification stage both in terms of hardware and software compatibility if the material is to be read by transportable.

9. In deciding if CBT is appropriate to a training problem, a value added analysis is essential to ensure that maximum and effective use is made of the capabilities offered by a computer based solution. If these capabilities are not fully utilised it is possible that an alternative, more cost effective solution to the training problem may exist.

10. The ability and background of the target trainees needs to be carefully considered in the presentation of the material (eg pace and comprehension levels) and the training strategy employed. Block use of tutorial CBT was found to be a poor strategy for basic training and the courseware was more effective when blended with traditional instruction. There is therefore a clear distinction between tutorial courseware designed for basic shore training and that which is produced for stand alone use at sea for continuation or refresher training.

COST/BENEFIT ANALYSIS

11. The remit for the evaluation required that cost benefit analysis be considered. For this training efficiency was defined as the relationship between the investment in training and the return on that investment. In theory, it should therefore be possible to compare the cost/benefit of a CBT solution with that of the traditional alternative. The evaluation considered that investment in this context could be examined in terms of both capital and running costs, whilst the return could be measured by the percentage of trainees who reach the pass standard and the time taken to achieve this. Some of the factors which have to be taken into account in each area are listed below.

- a. Capital Costs. Total initial investment, taking taking into account: other applications of the hardware and software which would reduce unit costs, investment of any Service expertise in the production of the material, system life expectancy (depreciation costs) and any initial costs of training staff to use the system.
- b. Running Costs. These include: maintenance of the training hardware and software, periodically recurring costs such as replacement of consumables and software modifications, licence fees, the training space required, marking costs, ongoing staff training costs, instructor costs and training support costs such as the provision of management data.
- c. Training Return. This was measured in terms of the percentage of trainees who reach the pass standard, offset against the overall time taken to achieve this (course length and any remedial training given). It is very difficult to cost other returns which may result from CBT such as: improved morale and motivation and increased consistency of both instruction and trained output. Moreover, to extend the scope of cost/benefit analysis to include training effectiveness involves the complexities of cost quantifying any improvements in operational job performance. In practice, the training return will not be known until the CBT solution is implemented. However, by comparing the investment costs for both the CBT and the traditional solutions it should be possible to assess the degree to which the training return must change if a cost benefit is to be achieved. Given this change it may then be possible, based on the experience of existing CBT applications, to assess if this is a realistic expectation.

In the event it was only possible to realistically cost/benefit analyse one of the projects and this showed that the CBT solution would, on maximum trainee throughput, pay for itself within 7 months despite its very high initial cost.

CONCLUSIONS

12. In drawing conclusions the study considered wider applications, attitudes and organisational impact to be sub-issues of acceptability, and impact on resources, job performance and instructional effectiveness to be sub-issues of training benefit. Clearly the weighting of these aspects may vary according to circumstances. For the purposes of this study the criteria for acceptability and training benefit were that, within the trial limitations, all the relevant sub-issues had to show a positive effect. Acknowledging the trial limitations the study reached the following conclusions:

- a. Tutorial CBT - Although it found acceptability in some applications, tutorial CBT in basic shore training was neither proved to be cost effective nor quantifiably better than existing training.
- b. Drill and Practice CBT - This was highly acceptable in the Naval training environment and was significantly better, both in terms of trainee performance and cost, than "traditional" training.

RECOMMENDATIONS

13. The following recommendations were made:

- a. CBT should only be funded where it can be proved to offer either the only solution to a training problem or where there are clear cost benefits.
- b. The potential of CBT should be trialled at sea for continuation and refresher training.
- c. Purely tutorial applications of CBT in basic shore training are not cost effective and should not be funded.
- d. Applications of CBT which involve drill and practice/emulation are highly effective and should receive funding.
- e. The use of CBT in simulation mode was not adequately addressed by the trial. Indications are that it could offer substantial benefits, and it is strongly recommended that a trial of simulation CBT be undertaken.

CBT DEVELOPMENTS POST EVALUATION

14. As a result of the recommendations of the evaluation report the policy for the development of existing CBT projects and the selection of new ones has been based on the following criteria:

- a. cost and training effectiveness.
- b. transferability to sea.

To ensure consistency in the application of this policy no funding is to be provided for any CBT project that had not been examined and approved by the RNSETT.

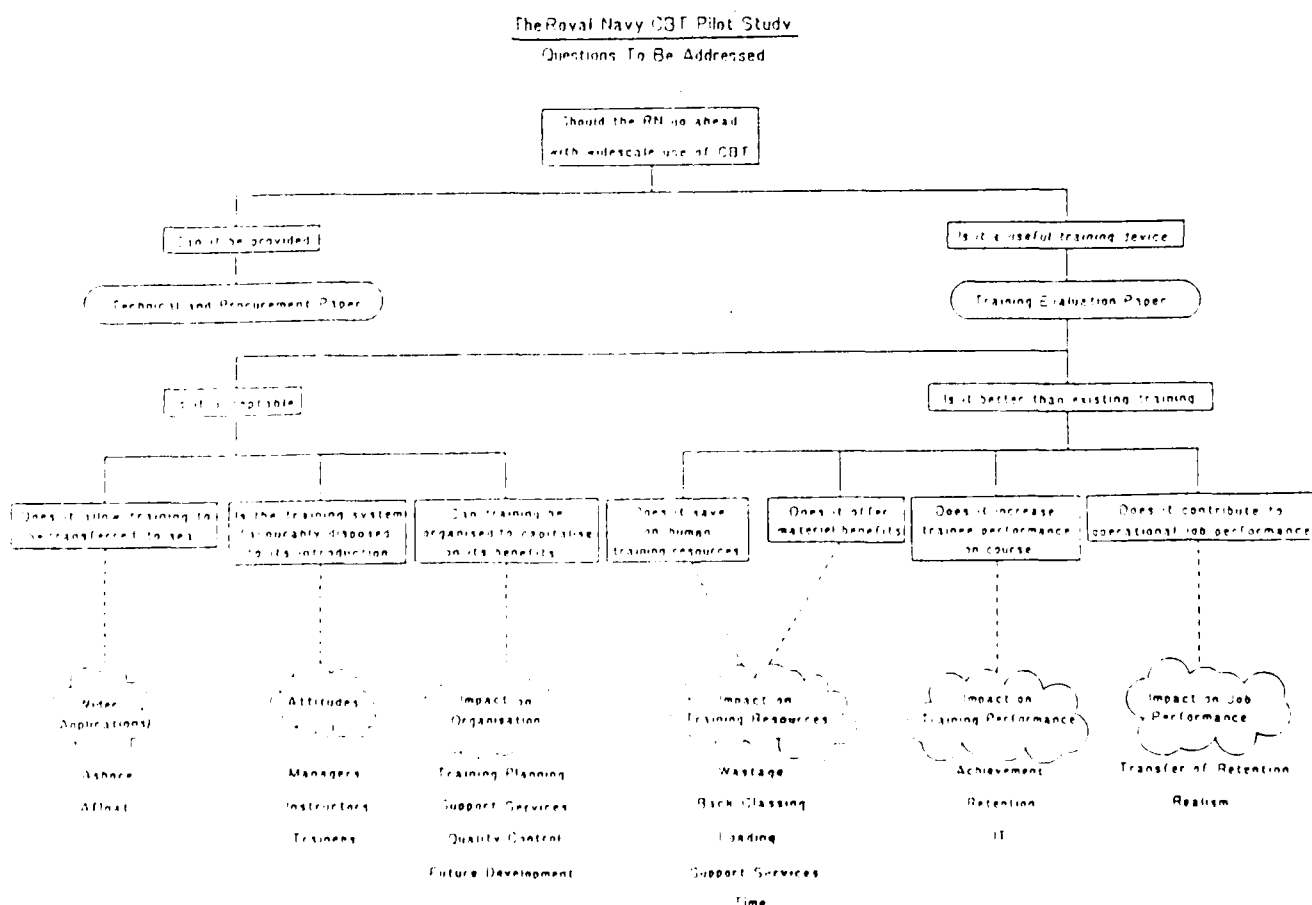
15. In accordance with the broad recommendations of the evaluation it was decided to concentrate the available funds on drill and practice training activities, procedural operator and maintainer training, and equipment emulation and simulation. As a result, the following projects were selected for development:

- a. Morse code transmission and reception training.
- b. Navigation in narrow channels and traffic separation schemes - simulation.
- c. Auxiliary Machinery Certificate (AMC) training - equipment emulation for operator and maintainer training.
- d. Nuclear propulsion plant emergency operating procedures - simulation.

Item a, b and c are envisaged to have applicability in at-sea continuation training whilst Item d will be used as a front-end trainer prior to full scale simulator training.

16. The RNSETT has prepared functional specifications for all 4 projects and will act as project managers from inception to installation. By doing this the school will provide both the training specialist expertise and the link between the end user and the contractor. Tenders have been invited from a wide range of UK CBT suppliers and the RNSETT is currently evaluating their responses prior to placement of contract. In addition to this the operational training vote is funding a series of CBT man-machine interface trainers for tactical, navigational and sonar equipment. The RNSETT is playing a lead role in the evaluation of appropriate solutions to these training problems.

17. As a result of the training evaluation, CBT in the Royal Navy is now being developed with the aim of achieving value for money investment in training by concentrating on cost effective drill and practice, emulation and simulation applications ashore, and investigating its use in these modes for at-sea continuation training.



TRAINING FOR A FLYING EMERGENCY

A TASK ANALYSIS OF ENGINE-OFF-LANDINGS

Amanda J W Feggetter, Heather M McIntyre,¹
Human Factors Unit, Headquarters Directorate Army Air Corps
Middle Wallop, Stockbridge, Hants. SO20 8DY. UK

INTRODUCTION

1. The UK Army Air Corps (AAC) expressed concern at the number of aircraft being damaged during the course of EOL training. In an 18 month period there had been 17 EOL incidents involving the Gazelle helicopter. As a result the Army Personnel Research Establishment's Human Factors Unit at the Headquarters Directorate Army Air Corps (APRE HFU HQDAAC) was tasked to undertake an objective study of EOL training.

2. In single engine helicopter training the ability to land the aircraft in the event of engine failure is taught at a very early stage. This exercise, called an Engine Off Landing (EOL) is associated with a high level of emotional content. Studies in human learning show that if skills are learned when the person is apprehensive or other strong emotional contexts prevail learning may be considerably impaired (Ernsting, 1978).

AIM

3. The aim of the study was to examine the reasons for the high number of training incidents which occurred during this particular exercise and to make appropriate recommendations for EOL training policy so that these might be avoided in future.

STUDY STRUCTURE

4. The study was carried out in two parts. A task analysis was undertaken to determine the skills needed to carry out an EOL. Flying training records were examined in terms of EOLs.

METHOD

TASK ANALYSIS

5. Interviews were conducted using the "Job Learning Analysis Questionnaire" (JLAQ) according to the standard method (Pearn M and Kandola R S 1987). Subjects included 2 aircrew from each of the following categories: experienced QHIs, less experienced QHIs, pilots, student pilots. The criterion for being included in the first category was to have more than 6 years instructional experience. Subjects were asked to give a brief description of the main aim or objective of an EOL. They were then asked to break down the task into its main activities. Each Main Activity was considered in turn and subjects asked to give a brief description of what was involved in this part of an EOL.

¹ Since this document was written personnel positions have changed as follows:
Amanda J W Feggetter, AGI(HF), Main Building, Ministry of Defence, Whitehall, London SW1.
Heather M McIntyre, Mentor, Computer Based Training, Bradford, Yorks.

6. The activities were then analysed in terms of 9 predetermined learning categories. These are as follows: physical skills, complex procedures, checking/assessing/discriminating, memorising information, ordering/prioritising/ planning, looking ahead/anticipating, diagnosing/analysing/solving, Interpreting written, pictorial, diagrammatic material, and adapting to new ideas/systems.

7. Subjects were asked whether these were relevant to each of the main activities they had identified. The pattern of responses was examined to determine the relative contribution of each learning category to the various stages of an EOL.

8. Throughout each interview relevant additional comments made by the interviewee about EOL training were noted.

TRAINING RECORDS

9. Subjects. The flying training records of 48 student pilots were included in the sample. A sub group of the above was entered into a microcomputer database for detailed analysis (Feggetter and McIntyre, 1985). The composition of the samples is shown below in Table 1.

TABLE 1: Sample Description

	<u>4 APC Courses</u>				Total
Whole Group	9	14	15	10	48
Sub Group	2	2	2	4	10

10. Analysis. A frequency count was made of the number of EOLs undertaken. This was related to the quality of EOL performance as assessed by flying instructors. An EOL rate per sortie was calculated for the instructors. Comments made by Instructors about students EOL performance were collated.

RESULTS

TASK ANALYSIS

11. Main aim of an EOL. All 8 subjects were consistent in their view of the main aim of practising EOLs. In the event of an engine failure the pilot must be able to carry out a safe landing. The primary aim was that the aircrew and passengers should survive with minimum injury. A subsidiary aim was to prevent damage to the aircraft. 62% of subjects also believed that practising EOLs built up student's confidence in their ability to cope with a real emergency.

12. EOL Breakdown. A definitive model of an EOL was derived from individual subject responses. This may be described in the following terms:

- a. STAGE 1: Briefing. This occurs both on the ground prior to the sortie and in the air to ensure there is no ambiguity between aircrew.
- b. STAGE 2: Preparation. This involves assessment of the landing area, weather conditions, and completion of checks. Choice of appropriate autorotative technique is followed by taking up the appropriate heading, height and speed. Lookout must be maintained throughout.
- c. STAGE 3: Autorotation. This involves lowering the collective lever, maintaining speed, eliminating drift and maintaining balance during

descent. Once certain the aircraft will reach the selected landing area the throttle lever is closed. The aircraft is then committed to land.

d. STAGE 4: Execution. There are two possible methods. The first, Variable Flare (VF), comprises the sequence flare, check, level and cushion. The second, Constant Attitude (CA), comprises the sequence check and cushion.

13. Figure 1 shows the generic EOL model. It can be seen that although there are variations in technique there is a central core of skills for all EOLs.

14. There are a number of decision points in the sequence. The first decision is made above 700 ft during the preparation stage. The pilot must select one of the 5 possible autorotative techniques. This decision is not bound by critical time constraints. The second decision is to enter autorotation at the appropriate height and speed. The third decision point is at 700 ft where the pilot must decide between the two possible types of landing VF or CA. The critical decision point lies between 150 and 100 feet when a further input from the pilot is required. If a VF landing is to be carried out the pilot must flare the aircraft at this point.

15. Analysis in terms of learning skills. The importance of the learning skills to each stage of an EOL are shown in rank order in Table 2 below. The briefing stage has been omitted.

TABLE 2: Rank ordering of skills by stage

Skills	Stage		
	Preparation	Autorotation	Execution
Physical skills	6	2	2.5
Complex procedures	4	7	5.5
Checking/assessing	1.5	2	2.5
Memorising facts	4	5	5.5
Ordering/planning	7	7	7
Looking ahead	1.5	2	2.5
Diagnosing/analysing	4	4	2.5
Using written material	8.5	9	9
Adapting to new ideas	8.5	7	8

Key: 1 = Most important.

- 9 = Least important

16. The findings show the importance of high level cognitive skills, such as checking/assessing and looking ahead, during the preparation and autorotation stages of the EOL. The cognitive skills used in the final stage are at a low level, mere visual monitoring and scanning. Physical skill is seen as relevant in all stages of an EOL, although less relevant in the preparation stage. It may be that when training pilots to carry out EOLs the specific skills to be acquired are those which occur early in the sequence, up to and including the final critical decision at 100-150 ft. The learning experience should therefore be concerned with high level decision-making skills. After this point relatively low level psychomotor skills come into play.

17. Miscellaneous Comments. Many comments were concerned with the degree of control given to the student during a practice EOL.

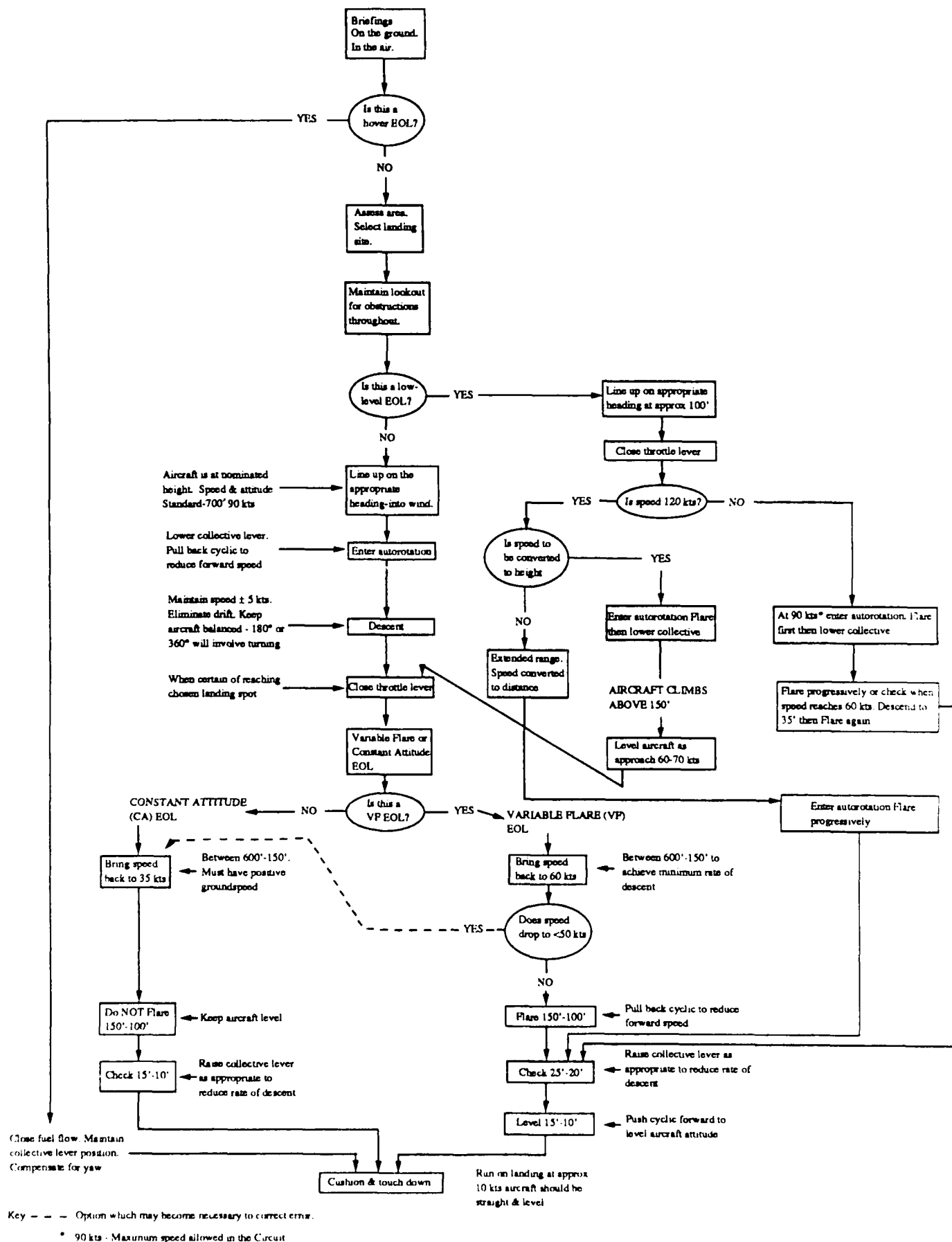


Figure 1: Procedure for a Practice Engine Off Landing (EOL)

TRAINING RECORDS

18. Table 3 illustrates the maximum, minimum and mean number of EOLs carried out. EOLs are practised during two phases of the course, Basic Rotary Wing (BRW) and Advanced Rotary Wing (ARW).

TABLE 3: EOL Statistics

Phase	Mean No of EOLs	Maximum No of EOLs	Minimum No of EOLs
All	85	118	59
BRW	57	88	29
ARW	28	47	14

19. Quality of EOLs related to number of EOLs. For each student the percentage of "acceptable EOLs and the percentage of "unacceptable" EOLs in each phase were determined. The calculations were based on instructor gradings of EOL performance. In the BRW phase there was a small correlation between student performance on EOLs and the number executed (Spearman rank correlation coefficient $p = 0.025$). Students with lower ability on EOLs undertake greater numbers in training. There was no significant correlation in the ARW phase. The number of EOLs is not related to student performance on EOLs.

20. Instructor Analysis. The ratio of number of EOLs per phase to the total number sorties flown in that phase was calculated for each instructor and illustrated in Table 4.

TABLE 4: EOL rates per sortie

Phase	Mean rate per sortie	Max rate per sortie	Min rate per sortie	Average no of sorties
BRW (N=13)	1.48	2.09	1.18	40
ARW (N=24)	0.75	1.50	0.28	56

There was a great variation between the extremes in both phases. In ARW instructors on average carry out 0.75 EOLs in each sortie. Six out of the 24 (25%) however have a rate greater than one per sortie, the average number of sorties in ARW being 56.

MICRO ANALYSIS ON TEN SUBJECTS

21. Coded comments. The frequency with which adverse comments were made by QHIs about different stages of EOLs are shown below (Table 5).

TABLE 5: EOL Comments by stage

Percentage of Comments	Phase			
	Preparation	Autorotation	Execution	Miscellaneous
N=254	15%	7%	52%	26%

52% of all the adverse comments made by instructors about EOLs are to do with the final execution phase.

DISCUSSIONS AND CONCLUSIONS

22. It is clear that throughout all stages of the EOL there is a requirement for physical skills. For the first three stages there is an additional need for high level cognitive skills whereas in the 4th and final stage the need is for physical skills alone. Effective training enables students to learn skills which can then be transferred to future similar situations. Physical skills which require little processing by the pilot are more readily transferred than the high level skills which require greater attentional involvement (Gibbs, 1970). The high level skills are specific to an EOL and are not related to any other element in the flying training course. The physical skills required throughout the procedure are practised in a number of other exercises during the course where transfer of learning can occur. This suggests that there should be more of an emphasis on the need to practice stages 1, 2 and 3 rather than on the need to practice stage 4, the execution of a landing.

23. Analysis of instructor comments reveal that it is stage 4 which receives the greatest number of critical remarks. This may reflect a bias on the part of instructors to be task orientated rather than skill orientated with undue emphasis being placed on the final completion of the EOL. It is during this stage that incidents occur and aircraft are damaged.

24. Skill acquisition. There is a wide variation in the numbers of EOLs practised by students, during both phases of the course. Although there is some slight relationship between ability at performing an EOL and the numbers executed in BRW such a relationship does not occur in ARW. All the training records examined belonged to individuals who had passed the flying course. Thus the individual who did the minimum number of EOLs must have reached a satisfactory standard of EOL performance.

26. Instructors. The study data suggests that much of the variability can be accounted for by the instructors. A number of factors noted during the JLAQ interviews may account for this variation. These are: a) Enjoyment - Instructors appear to like showing off their skill. b) Convenience - An EOL at the end of a sortie appears to provide the quickest way to the airfield. c) Continuation training - Instructors have said that student training sorties provide them with their only opportunity for their own continuation training. d) Image - Instructors wish to maintain a "macho" image in front of their students. e) Mystique - Instructors surround the exercise with an air of mystique.

REFERENCES

- Downs, S., 1986: Unpublished MOD visit report to RNAS Culdrose.
 Ernsting, J., 1978: Aviation Medicine. Tri Med Books.
 Feggetter, A. and McIntyre, H., 1985: A Microcomputer Based Database Processing System to Investigate Human Factor Aspects of Aircraft Accidents and Incidents. Proceedings of 26th Annual Conference of the Military Testing Association, Munich.
 Gibbs, C.B., 1970: Servo Control Systems in Organisms and the Transfer of Skill. Skills Ed by Legg D. Penguin.
 Pearn, M. and Kandola R.S.: Job Learning Analysis: User Guidelines. Oxford.
 Pearn, M. and Kandola, R.S., 1987: Case Studies using the JLA. Sheffield: MSC

HELICOPTER INSTRUMENT PROCEDURES TRAINER (HIPT)
AN EVALUATION OF TRAINING EFFECTS

CAPTAIN I.E. (ED) WIEBE
CAPTAIN G.A. (GREG) REISER

CANADIAN FORCES CENTRAL FLYING SCHOOL
WINNIPEG, MANITOBA R3J 0T0

HELICOPTER INSTRUMENT PROCEDURES TRAINER AN EVALUATION OF TRAINING EFFECTS

BACKGROUND

1. Prior to 1979 initial pilot training in Canada consisted of a 200 hour "wings course" followed by specialized training on various aircraft types. Pilots selected for helicopter training from this program received the necessary instruction in instrument flight to make the transition to helicopters. A decision to "stream" fixed wing and helicopter candidates at a pre-determined 140 hour cut-off was cause for concern in the helicopter training environment regarding instrument standards. This decision, combined with the introduction in 1981 of the Jet Ranger III with its expanded instrument package, further compounded instrument standard concerns. Eventually, as could be expected, the Operational Training Units (OTUs) noted a decrease in instrument flying proficiency of graduate helicopter pilots.

2. A requirement existed for an expanded syllabus for the Basic Helicopter Training Course (BHT). Significant course content changes and new procedures would have to be introduced into basic helicopter training to upgrade the proficiency of the BHT graduate. The course syllabus was rearranged, however, no increase in course time for instrument training was allotted. Therefore, to permit flying instruction to the full scope of the revised syllabus, instructors had to pare concentration on some aspects of the revised instrument phase.

3. The acquisition of a Helicopter Instrument Procedures Trainer (HIPT) seemed the logical solution to offset the limited actual aircraft hours and complement air instruction to a degree which would improve the overall proficiency of BHT graduates.

4. To place this acquisition in perspective there are 380 designated helicopter cockpit positions in the Canadian Forces (CF). 3 Canadian Forces Flying Training School (3 CFFTS) is currently tasked to conduct 5 courses annually designed to produce 80 graduates.

AIM

5. The aim of this paper is to provide an overview of the effect the HIPT has had on ab initio helicopter training. This will include:

- a. a review of graduate assessment by OTUs;
- b. an introduction to the HIPT basic system;
- c. a review of the HIPT instrument training syllabus;

- d. an internal evaluation of HIPT training;
- e. a comparison of graduate Instrument Rating Tests (IRTs); and
- f. a review of maintenance benefits.

REVIEW OF GRADUATE ASSESSMENT

6. An independant pilot survey conducted by the Canadian Forces recommended the "streaming" of fixed and rotary wing student pilots following 140 hours of instruction on a basic jet trainer. The introduction of the "Wings" course for helicopter pilots in 1979, with its subsequent loss of 24.6 hours of applied instrument instruction, drastically reduced the quality of instrument proficiency to that of previous "wings" graduates.

7. After receiving the graduates (post 1979), the OTUs were quick to identify weak areas of instrument proficiency. Specifically:

- a. basic instrument and flight procedures;
- b. in-flight awareness (air picture); and
- c. IFR enroute including Radio Telephony (R/T) procedures.

8. Because of these noted deficiencies, the training syllabus was revised to include a total of 6.6 hours dual air instruction plus an additional 2.0 hours mutual. This instructional time concentrated on applied instrument procedures designed to increase the students situational awareness and to hone basic instrument skills. Although the OTUs recognized the increased benefits derived from this additional instruction, their considered opinion remained, that graduates lacked the required basic instrument flight skills and instrument procedural knowledge to successfully upgrade new graduates to aircraft captaincy status.

9. To further address the loss of applied instrument instruction for BHT candidates, the decision was made to introduce an instrument procedures trainer (HIPT) to fill this training void.

THE HIPT BASIC SYSTEM

10. Physically, the HIPT features an accurate, high fidelity crew station with actual simulated aircraft instrumentation. To maximize transfer of training, the system provides for:

- a. dynamic helicopter systems simulation with instructor controlled fault introduction;
- b. full instrumentation flight rules (IFR) capability covering all phases of instrument flight;
- c. simulated map area representing real world navigation;

- d. audio simulation of engine, rotor and systems warnings;
- e. full weather simulation;
- f. automated lesson plan capability; and
- g. play back feature.

11. The instructor's station is connected directly to the simulator crew station to permit easy student monitoring and instruction during exercises. All simulator control is affected through the instructor's terminal via user friendly keys and menu driven displays. The instructor is able to freeze, re-start or re-initialize the system at any time and use the "jump" feature to reduce the length of tedious cross country legs and maximize training tasks.

THE HIPT INSTRUMENT TRAINING SYLLABUS

12. The intent of the HIPT and the proposed training syllabus was designed to familiarize the BHT student with the Jet Ranger cockpit and to introduce applied instrument procedures prior to the corresponding air lessons. The training syllabus emphasized the following:

- a. Basic starting/shut down procedures
 - hot start procedures
- b. Emergencies
 - all ground and air related malfunctions
- c. Basic instrument flight
- d. Applied instrument procedures
 - Automatic Direction Finder (ADF) procedures
 - Omnidirectional Range (VOR)
 - Precision Approach Radar (PAR)
 - Enroute Procedures

INTERNAL EVALUATION OF HIPT

13. Nine student courses beginning early 1986 have completed the revised BHT course utilizing the HIPT. An internal evaluation conducted with instructional staff and students was overwhelmingly supportive of the benefits derived from HIPT training.

14. Student productivity increased significantly as evidenced by a renewed self confident approach and performance in the air. More specifically, evaluation results determined:

- a. that students are more proficient at cockpit checks and emergency procedures;
- b. that instructor pre-flight briefing times are reduced; and
- c. that students demonstrated an increase in procedural knowledge and basic instrument skills.

A COMPARISON OF GRADUATE IRT PROFILES

15. The Canadian Forces administers annually an Instrument Rating Test (IRT) to all military pilots. The test consists of instrument flying skills or tasks that must be performed by the pilot and graded by an evaluator on a proficiency level basis. An overall assessment on the performance of all tasks is then made which ultimately results in the successful pilot being awarded an instrument rating ticket. For students undergoing helicopter training the IRT is the culmination of their training in the instrument phase.

16. A comparison of IRT profiles of 90 pre-HIPT and 85 post-HIPT graduates indicates a fairly significant increase in performance proficiency of post-HIPT graduates. Using the following performance rating criteria, results are as indicated in Figure 1.

- Superior** - performed all tasks consistently better than standard
- Above Average** - performed majority of tasks better than standard
- High Average** - performed some tasks better than standard
- Average** - performed all tasks to standard
- Below Average** - experienced some difficulty but met standard
- Unsatisfactory** - failed to meet standard

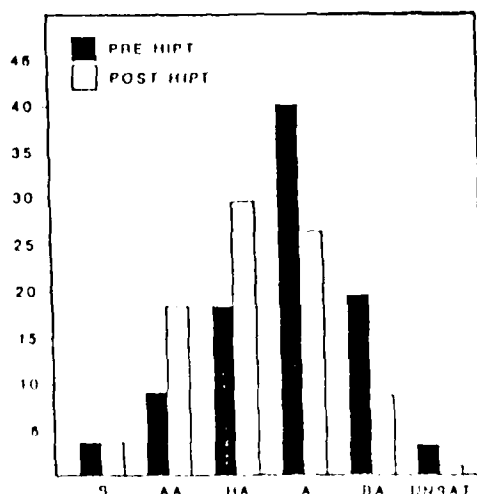


FIGURE 1

A REVIEW OF MAINTENANCE BENEFITS

17. A review of the CH 139 maintenance program has provided some interesting results. Over a four year sampling of BHT courses, a total of nine Hot Starts were recorded, all in the two years preceeding the HIPT. Each hot start occurrence and its resultant maintenance costs is represented by the following statistics:

5 Hot Start Inspections with no resultant damage

Inspection of 10 man hours per aircraft
Downtime minimal

4 Hot Start Inspections with resultant damage

Inspection of 50 man hours per overhaul
Engine overhaul @ \$30,000 each
Downtime of 3 months per overhaul

18. The decrease in the number of Engine Hot Starts since 1986 is contributed solely to the increased attention to training in the HIPT.

CONCLUSION

19. The HIPT has proven to be an invaluable acquisition for the CF. Student productivity/proficiency in all aspects of instrument flying, combined with an increased student self confidence, has resulted in the graduation of a more competent helicopter pilot. In addition, 3 CFETS has realized considerable savings in terms of man hours, overhaul/replacement costs and downtime for the Jet Ranger III.

PERFORMANCE EVALUATION IN TACTICAL TRAINERS

John A. Modrick and Thomas A. Plocher

System and Research Center, Honeywell Inc.

Our goal was to develop a performance assessment system (PAS) as an instructor's aid for evaluating operators, decision makers and teams executing tactical actions and decisions in simulated exercises on tactical trainers, emphasizing cognitive activities. The PAS enhances the instructor's capability to monitor, describe, understand and evaluate the behavior of individuals and teams and thus provide feedback and critique appropriate to the behavioral objectives of an exercise. It carries out four functions: track behavior, evaluate the responses against criteria, recognize and report errors and compile information for After Action Review (AAH) and other performance records. It reduces instructor workload, enabling him to monitor several teams concurrently and devote more attention to management of instruction.

This work was internally funded as Independent Research and Development.

APPROACH TO A PERFORMANCE ASSESSMENT METHODOLOGY

We developed a procedure for computer monitoring and evaluation of the performance of tactical procedures. We used the concepts of semi-structured tasks and script to devise a representation of the behaviors and knowledge in tactical procedures. A script of a tactical activity is a sequence of actions and events necessary to achieve a tactical goal. Criteria and errors for the behaviors are embedded as frames in the scripted action steps. We used the script to 1) structure tactical actions as procedures, 2) integrate associated knowledge and skill into the procedures, 3) embed performance criteria into the procedures and 4) infer which procedural script the individual was executing. The script sequence is also one of the performance criteria. A diagram of the principal functions in the method is presented in Figure 1.

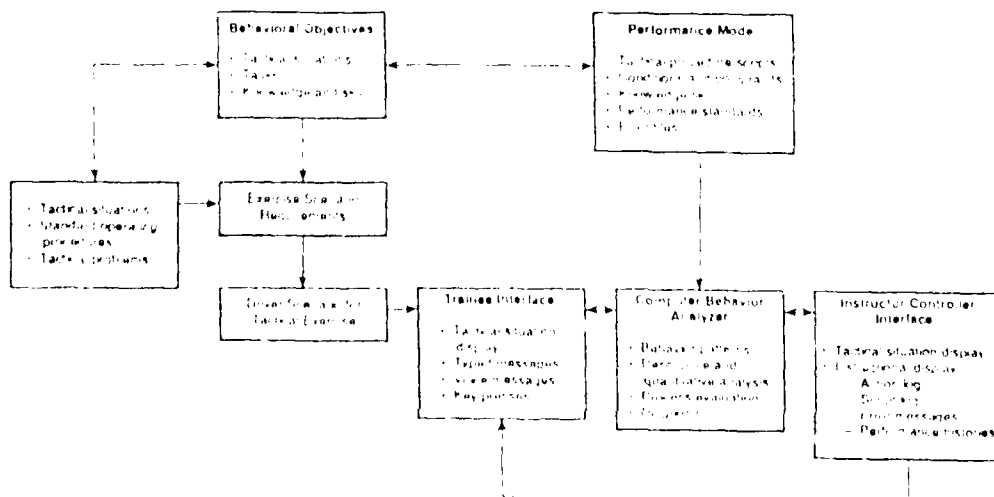


Figure 1. Principal Functions of a Performance Assessment System.

The objective of human performance measurement is to determine mastery of tasks, procedures and underlying units of knowledge or skill. Typical conventional measures are a product measure of how much was accomplished during a time period. Evaluation is served better by process measures which deal with what actions were executed, their appropriateness and how well they were executed. They are more applicable to complex operating procedures which are flexible, goal-directed response sequences whose configuration is determined by situational conditions and constraints but can be reconfigured to adapt the response sequence to situational conditions.

DEVELOPMENT OF THE PAS METHODOLOGY

Developing such a system depends on solving three technical problems: Representation of complex, multi-alternative, tactical activities in dynamic situations which unfold over time; computer recognition of operationally meaningful patterns of tactical behaviors; and computer-aided assessment of performance against quantitative standards and diagnosis of errors.

Representation of Tactical Activities

We developed procedural scripts and performance measures in two areas: search procedures for anti-submarine warfare (ASW) and console operating procedures for an airborne target acquisition system (Joint STARS). The ASW scripts, based on Crowe et al. (1981), were activities of an ASW team on a destroyer in a Carrier Battle Group. The scripts prepared are Sprint to a New Position, Drift and Listen and Lay Sonobuoy Barrier. They include coordinated action and communication among several individuals (ASW Operations Coordinator, Anti Submarine Air Controller, LAMPS helicopter pilot, Underwater Battery and the Battle Group's ASW Coordinator). The function of Joint STARS is analysis of imagery for target acquisition in a moving target indicator radar system. The scripts implemented, based on Plocher et al (1985), are Search For Target Candidates, Track Target, Update Target and Report Targets. They involve a two person team: Search and Tracking Operator and Tracking and Surveillance Supervisor.

Computer Recognition of Behavior Patterns

A procedure was formulated for computer recognition of patterns among detected responses events (Implementing Steps) that correspond to actions from the scripts. These responses events are key presses, messages and displayed information. The basic performance data are a continuous, temporally ordered listing of sensed responses events. The lowest level of performance measurement is describing the behaviors actually executed. Responses events are clustered into patterns which correspond to actions in a procedural script. An identified action is used to identify candidate procedural scripts which the operator might be executing. The most probable procedural script is reported to the instructor as the activity being executed. The data and reasoning on which this conclusion is based are also reported.

Quantitative Evaluation of Performance

Quantitative performance evaluation requires standards of response times or errors against which observed behavior can be compared. The sources of standards are performance records of comparable operators and "a priori" values for acceptable performance based on doctrine. The first level of evaluation is based on the continuous, temporally ordered record of responses and events on a console. They are compared to quantitative criteria for response times and errors. Error messages are sent to the instructor. Out-of-order and extraneous responses are identified and referred to error analysis and reporting. Further evaluation deals with the tactical appropriateness of the behavior executed, analysis of errors and generation of feedback to instructor and operator. A major development goal is collation of performance data, critical incidents and events from the exercise into a summary of performance for an AAR.

DEMONSTRATION OF THE PAS METHODOLOGY

A preliminary "proof of principle" demonstration was developed (Parrish et al. 1988) to provide evidence that the scripts could be structured and hosted on a computer and performance assessment can be done. The scripts from ASW and Joint STARS were implemented in "bare bones" driver scenarios for simulated tactical actions. The ASW demonstration had a minimal simulation capability and was hosted on an IBM XT; the driver scenario consisted of a sequence of static situational displays. The demonstration system for Joint STARS consists of a Microvax 11780 computer and a Silicon Graphics display from the Joint STARS Ground Station Simulator augmented by an IBM XT used as an Instructor/Controller console. The more extensive computer capability was used to display actual target imagery taped from field exercises.

THE SCRIPT-LIKE CHARACTER OF TACTICAL ACTIVITIES

Our method for representing tactical activities and the knowledge required for executing them is based on the concepts of and semi-structured task (Keene and Morton, 1978) and script (Schank and Abelson, 1977). Cognitive activities such as tactical decision tasks typically are semi-structured activities in emergent situations which unfold as one is performing the tasks. It is impossible to preplan them in any detail or treat them as linear, invariant sequences. The actor in these situations must be able to structure incoming information, recognize the situation, adapt his perception as it unfolds, choose an action and modify it with significant changes in his perception of the tactical situation.

The structure and properties of a script and semi-structured task are a conceptual framework for representing the complex tactical activities of individuals and teams in a goal-oriented hierarchical structure of actions, subgoals and goal state. The steps in a tactical script are action states, each consisting of a subgoal, actions necessary to transition between subgoals, options available for subsequent action states and the conditions and constraints under which these transitions can be made. They are best represented as trees or a hierarchical structure of nodes. Each node is an action state.

A tactical script is an analogue to the standard operating procedure. Situational and sequential uncertainty in emergent situations and semi-structured tasks are handled in the operational world by classifying a situation in terms of a set of standard situations or schema and reacting to it according to an associated standard sequence of actions. A procedural script has a standardized set of initial conditions and actions similar to opening moves in chess. The nominal procedure must be modified in execution, however, to adjust to local tactical conditions, enemy actions and outcomes. Therefore, the representation of standard tactical procedures must provide a sequential structure with branching and include the conditions, contingencies and constraints that relate to selection among options at each branch point.

Many military manuals, particularly the U. S. Army's how-to-fight manuals, are procedure oriented with a strong script flavor. They can be cast readily into a script format but must be expanded to embrace the complexity of standard procedures in operational environments. They must be augmented with implicit procedural knowledge and explicit domain knowledge which are part of the lore of the tactical community. However, this knowledge and the performance standards are seldom stated in any documents or even well articulated.

SCRIPT AS A PSYCHOLOGICAL CONSTRUCT

The script as a psychological construct has been the object of research as a mechanism for memory and computer understanding of text such as newspaper stories (Schank, 1979). Memory research is the recall of events or activities extended over time such as eating in a restaurant or getting up of the morning. Recent work summarizing extension of memory and understanding into areas such as explanation, argument and persuasion are summarized in Galambos et al. (1986). Research on computer understanding can be represented by a computer "reading" newspaper articles about a class of phenomena such as natural disasters (earthquakes) and reporting summaries of, or answering questions about, the events (Cullingford, 1978). Constraints and relationships in a script resolve ambiguity and reduce the parsing for natural language understanding.

Scripts are the mental models or schemata by which humans organize their knowledge about situations, events and activities that extend over time; they function to organize information, recognize situations, select actions and guide their execution. They influence how complex activities are learned and recalled and have implications for instructional strategy. The script model determines how events are perceived, alternative actions are recalled, selections are made among alternatives and action is planned and controlled. The model of semi-structured-tasks script then could be a good framework for performance aiding in complex, cognitive tasks.

Our approach assumes that a person has a repertoire of basic, procedural action skills which are chained sequentially and combined in parallel time sharing to form more complex tasks and activities. They are learned in basic combat training and the early phases of job specific training. Supporting knowledge becomes integrated into the procedures as mastery proceeds. The interconnections among these skill/knowledge structures becomes more complex and intricate with experience.

Recent research suggests the idea of a script complex: the combination of a standard tactical situation, a set of situational variants and a hierarchy of scripted procedures associated with each situation and variant. It combines the situational schema (Noble and Grosz, 1986) and the combination of perceived situation and action alternatives from recognition-based decision making in fire ground commanders (Klein et al. 1986). Alternative procedures are associated with tactical situations ordered in priority. The alternatives are coded by the conditions and constraints under which they are applicable.

DISCUSSION

This approach to performance measurement is feasible. Further, the research has led to better definition of the problems and terminology and identification of problems and issues which must be addressed.

Refinement of Script Terminology

We had to invent and refine some terminology. First, we adopted the term *Structured Procedure* to differentiate procedural scripts from the fixed procedures in maintenance tasks. Second, we refined the term "tactical" by differentiating three kinds of scripts:

1. **Tactical Procedure Scripts:** allocate, deploy and maneuver resources (platforms, sensors and weapons).
2. **Operating Procedure Scripts:** control-display actions for operating a system or console in a tactical or operational environment. Tactical factors are seldom considered explicitly beyond the initial deployment and set up of the system; the operator typically is busy responding in real time to immediate threat conditions.
3. **Analytical Procedure Scripts:** analysis and interpretation of tactical events.

We distinguished two phases of performance measurement: descriptive and evaluative. Descriptive measurement is sensing and describing the behaviors actually executed. Evaluative performance measurement deals with the appropriateness of the observed behaviors, analysis of errors and generation of feedback to the examiner and operator. Feedback can take several forms such as: Right/ Wrong, Corrective, Explanation for Errors and AAR. A summary of operator performance on an exercise is compiled as an adjunct to the After Action Review. The final category of evaluation is diagnosis of deficiencies in the knowledge or skill.

Performance measurement must be tied to explicitly stated behavioral objectives, a statement of the task, procedure or unit of knowledge or skill on which the operator is to be examined. They must be stated as acceptable action sequences, what to observe, and standard tactical situations and conditions under which they are appropriate. The information must be translated into specifications of a simulation scenario to be executed.

SIGNIFICANT ISSUES TO BE ADDRESSED

1. Feasibility of executing in near real time the computation required for script-based performance assessment in a full scale simulation exercise.
2. Acquiring the situational and procedural knowledge about the tactical domain and performance standards from qualified members of the tactical community.
3. An authoring system to provide the tools with which to write and update software, implement and support the functions of performance assessment and prepare the courseware for the simulated exercises.
4. An exercise compiler to enable the instructor to compile exercises which will satisfy the conditions necessary to evaluate a player's mastery of a training objective.

REFERENCES

- Crowe, W. Hicklin, M. K. Obermayer, R. and Satzer, W. Team Training Through Communication Control. NAVTRAEQUIPCEN 80-C-0095-2, Naval Training Equipment Center, Orlando, FA 32813.
- Cullingford, R.E. Script Application: Computer Understanding of Newspaper Stories. Yale University Research Report 116. Office of Naval Research, Arlington, VA. 1978. (AD A056 080)
- Galambos, J. A., Abelson, R. P. and Black, J. B. (Eds.) Knowledge Structures. Lawrence Erlbaum Associates, Publishers. Hillsdale, NJ. 1986.
- Keen, P.G. and Morton, M.S.S. Decision Support Systems: An Organizational Perspective. Addison-Wesley Publishing Co., Reading, MA. 1978.
- Klein, G. Calderwood, R. and Crandall, B. Recognition-Primed Decision Making. Fourth Annual Workshop on Command and Control Decision Aiding, Kansas City, MO, 1986.
- Noble, D. and Grosz, C. Schema-Based Decision Making. Fourth Annual Workshop on Command and Control Decision Aiding, Kansas City, MO, 1986.
- Parrish, E. J. Plocher, T. A. Modrick, J. A and McGuigan, S. M. Performance Measurement in Tactical Team Training. Independent Research and Development Project No. 130603008, Honeywell, Inc. Systems and Research Center, Minneapolis, MN, 1988.
- Plocher, T. A., Tamanaha, R. and Gilles, P. METACREW: A Simulation of Joint STARS Crew Resources. Technical Report, U.S. Army CECOM, Ft. Monmouth, NJ, 1985.
- Schank, R. C. Reminding and Memory Organization: An Introduction to MOPS. Research Report 170, Yale University, Department of Computer Science, New Haven, CT. DARPA/ONR Contract N0014-75-C-1111, 1979.
- Schank, R. C. and Abelson, R. Scripts, Plans, Goals, and Understanding. Hillsdale, NJ: Erlbaum, 1977.
- 21 November 1988

EFFECTIVENESS OF AN ADVANCED INDIVIDUAL

COMBAT ARMS TRAINER (AICAT)

Brigadier General Charles R. White
Headquarters Air Training Command
Randolph Air Force Base, Texas

Captain Jarean L. Carson
Headquarters Air Force Military Personnel Center
Randolph Air Force Base, Texas

Captain Keith R. Wynkoop
Master Sergeant Edward B. Scott
Air Force Military Training Center
Lackland Air Force Base, Texas

First Lieutenant Von M. Cameron
USAF Occupational Measurement Center
Randolph Air Force Base, Texas

Clifford A. Butzin, Ph.D.
Wilford Hall USAF Medical Center
Lackland Air Force Base, Texas

Abstract

New advances in technology make increasingly capable simulation available at decreasing costs. In marksmanship training, simulators may improve training quality while reducing training expenditures. Previous research indicates that simulators may be at least as effective as traditional methods of instruction. Currently, the Air Force Combat Arms School at Lackland Air Force Base is testing an Advanced Individual Combat Arms Trainer (AICAT) which simulates both the M-16 rifle and the M-9 pistol. This study examines the effects of noise, recoil, and training time on M-9 qualification scores of 166 security police trainees. Results indicate the AICAT is effective for training marksmanship skills, and consistent recoil and report are important to the training process. Future research should pursue application to both training and initial screening for training.

Effectiveness of an Advanced Individual

Combat Arms Trainer (AICAT)

Previous research has shown small arms simulators to be effective in training while saving valuable ammunition. Much of the research has concentrated on effectiveness of a particular device that simulates the M-16 rifle, and not necessarily on simulation of a handgun or on how the device is used. One research study compared several different M-16 simulators to traditional training, showing that part-task trainers can be at least as effective, or in some cases more effective, than traditional training methods (Mehlenbeck & Collins, 1986).

Several other studies have examined the training effectiveness of the Joint-Service Multipurpose Arms Combat Simulator (JMACS), a part-task trainer that uses a Commodore 64 personal computer and a demilitarized M-16 rifle. Results from these studies have been somewhat inconclusive. One study performed by Perkins, Selby, Broom, and Osborne (cited in Schroeder, 1987) showed no significant differences in live-fire performance between control and experimental (JMACS training) groups. Another study performed at Lackland Air Force Base (AFB) showed trainees who used JMACS achieved slightly higher scores than control groups, though the differences were not significant (Eagle Technology, Inc., 1987).

The issue of how the simulator is used relates to the amount of time a trainee spends with the device and to its fidelity. According to Schneider (reported in Hogan, Arneson, and Salas, 1987), the optimum amount of practice varies by individual characteristics and the training situation. Fidelity may also vary by individual characteristics. Buffardi and Allen (reported in Hogan et al., 1987) conclude:

Simulators high in physical ability seem to aid low ability subjects to a greater degree than high ability subjects, whereas a more cognitive schema (low physical fidelity) facilitates the performance of those with higher abilities more than those with low ability levels.

This study examines the effectiveness of an M-9 handgun simulator in training marksmanship skills, and also looks at the importance of recoil and report (simulator fidelity) and amount of time it is used.

Method

Subjects

One hundred sixty-six male and female trainees in the Law Enforcement (LE) technical training course for the Security Police career field served as subjects for this study.

Apparatus

Advanced Individual Combat Arms Trainer (AICAT). AICAT is an interactive videodisc (IVD)-based system that simulates range fire for various small arms. This study used a system consisting of eight stations, each configured with a demilitarized M-9 handgun, a tricolor projection unit, and an 8-foot by 8-foot projection screen. Compressed CO₂ gas provided simulated recoil. Digitally recorded sound provided weapon report.

Design and Procedures

Trainees were randomly assigned to a traditional 3-day weapon qualification block in a 33-day LE specialist course and divided into a control group and treatment groups. All received instruction on M-9 handgun nomenclature, safety, maintenance, and dry-fire shooting. The control group received 30 minutes of practice using pencil shot drills and triangulation techniques. In place of this practice, the treatment groups used AICAT in one of eight conditions varying by recoil (no recoil/full recoil), report (no report/full report), and time (10 minutes/20 minutes). Next, all trainees fired on the range according to the Air Force Qualification Course (AFQC) of fire. The AFQC requires a trainee to perform one practice order of fire before an evaluation order of fire. Trainees fired at a 9 1/2-inch by 40-inch male "E" kneeling target, with a 10-inch center ring, at 7-yard, 15-yard, and 25-yard distances. Instructors recorded the total number of hits on silhouette and total number of hits in the 10-inch center ring for practice and evaluation orders of fire for each trainee.

Results

The statistical analysis used Analysis of Variance (ANOVA) to look at overall differences (comparing the control group to the total AICAT group) and differences in the AICAT conditions (recoil/no recoil, report/no report, 10 minutes/20 minutes).

Table 1 shows the mean scores of those who used AICAT and the control group. The mean score for those who used AICAT was higher in every case (number of hits on center, practice fire; total number of hits on target, practice fire; number of hits on center, evaluation fire; total number of hits on

target, evaluation fire). Analysis showed the difference was statistically significant for number of hits on center for practice fire ($p < .05$).

Table 1

Overall Comparison of Mean Firing Scores

<u>Measure</u>	<u>AICAT</u>	<u>Control</u>
Practice Fire		
Hits on Center*	40.38	36.50
Hits on Target	58.67	57.80
Evaluation Fire		
Hits on Center	20.53	19.56
Hits on Target	33.81	33.38

* $p < .05$

Second, a look at the varying conditions for those who used AICAT showed statistically significant interactions between recoil and report on the evaluation fire, both for hits on center and for total number of hits on target. The groups who had both recoil and report or neither recoil nor report had significantly higher mean scores than the groups who had one and not the other ($p < .05$ for hits on center, $p < .058$ for total number of hits on target). Analysis showed a statistically significant 3-way interaction between recoil, report, and time ($p < .05$ for both hits on center and total hits on target for evaluation fire). Examining this interaction showed the recoil/report effect described above is predominant for those who used AICAT for 20 minutes versus 10 minutes (see Table 2). Analysis showed no significant effects of recoil, report, or time on mean scores for practice fire.

Table 2

Recoil/Report/Time Interaction in Evaluation Fire

Time	Report			
	yes		no	
	Recoil		Recoil	
	yes	no	yes	no
Mean Number of Hits on Center*				
10 min	19.40	19.33	20.20	20.30
20 min	24.50	16.00	19.50	23.25
Mean Number of Hits on Target**				
10 min	33.70	33.67	34.60	33.80
20 min	34.86	32.36	32.70	34.42

* $p < .05$. ** $p < .05$

Discussion

The AICAT seems to be an effective training device, though additional research is necessary. For the practice fire, AICAT seems to be more effective than the traditional methods of instruction that it replaced, at least in training marksmanship skills reflected in the finer distinction of hits on center (versus total number of hits on target). While the second fire for evaluation showed no significant differences, the mean for those who used AICAT was slightly higher, indicating that AICAT is at least as effective as traditional methods. Considering the varying conditions (report, recoil, time) for using AICAT, the strongest use for evaluation fire seems to be consistent report and recoil for a duration of 20 minutes. Further analysis comparing the control group to this configuration may show a significant effect of AICAT on evaluation fire, as well.

References

- Eagle Technology, Inc. (1987). Training effectiveness evaluation of the Joint-Service Multipurpose Arms Combat Simulator (JMACS) in the United States Air Force Marksmanship Instructor and Security Police Law Enforcement Courses: Summary report of statistical data (Contract No. F41689-86-D-009). San Antonio, TX: Author.
- Hogan, Arneson, and Salas (1987). Individual Differences in Military Training Environments: Four Areas of Research (NAVTRASYSCEN No. TR 87-003). Orlando, FL: Naval Training System Center. (DTIC No. AD-A187 463).
- Mehlenbeck, C.W. & Collins, K.G. (1986). Concept evaluation program (CEP) test: M-16 rifle Gowen South phase I (USIB Project No. 3782). Fort Benning, GA: Training and Doctrine Command, US Army Infantry Board, Small Arms Test Division.
- Schroeder, J.E. (1987). Overview of the development and testing of a low-cost, part-task weapon trainer. Proceedings of the 1987 Conference on Technology in Training and Education, pp. 200-209.

Joint Service Training Requirements Decision Support System

Dr H. Barbara Sorensen
Ms Marlene R. Laskowski
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas 78235-5601

Introduction

The on-schedule deployment, effective maintainability, and supportability of all Department of Defense operational weapon systems all dependent on the efficient and timely training of soldiers, sailors, marines, and airmen that maintain and operate these systems. The joint services need to ensure the development, testing, and installation of training systems in time to support the deployment of emerging weapon systems. The official interservice procedure for developing this training system is the Interservice Training and Review Organization's Instructional Systems Development (ISD) process. An ISD process is required for all training development programs, including those for equipment in the acquisition cycle.

The current ISD process for major weapon system acquisitions encompasses the collection and analysis of Logistic Support Analysis Record (LSAR) data, engineering drawings and specifications, and the decisions made by experienced training analysts, management personnel, design engineers, and system users. All major weapon system acquisitions require both ISD and LSAR to be performed throughout the system life cycle.

Automating the entire process for all three services will save substantial time and resources to ensure trained maintenance personnel will be in place when the weapon system is fielded. Currently, variations of the ISD process are applied by each service producing delays and confusion for contractors. In addition, monitoring activities and quality control efforts become difficult. Analysis of the joint service ISD process and the information support requirements needed from LSAR to speed the LSAR-to-ISD process and achieve efficient, standard procedures is required.

Development of a Training Requirements Decision Support System will apply an information modelling approach to identify the tri-service information requirements for ISD and assess the feasibility of a tri-service computerized decision support system incorporating an LSAR-to-ISD interface.

The Joint Service Training Requirements DSS will be a powerful tool for performing ISD analyses of weapon systems. The automated LSAR-to-ISD data interface will support front-end analyses of training system requirements. The LSAR interface should improve the quality of information exchanges between ISD analysts and the weapon system design engineers and result in a wider range of training issues being addressed in a more complete fashion. Both the early analysis of training requirements and the interaction with the system designers will contribute to the development of more effective training systems.

In addition to the LSAR data interface, the Joint Service Training Requirements DSS will provide decision support to the training system designer, specifically to permit more efficient analyses of alternative training approaches. The decision support and data management features will enable the training system analyst to work more productively. The Training Requirements DSS is expected to demonstrate sizable quantitative and qualitative improvements over current procedures for conducting, managing, and effectively using the products of ISD analyses.

Joint Service Training System Requirements

Instructional Systems Development (ISD) is the systems engineering approach to training. ISD consists of a structured series of analytical steps that break down a weapon system's operational, maintenance, and support requirements into specific tasks, task elements, skills, and knowledges. ISD considers the relative need and appropriate method to train each task, task element, skill, and knowledge to a target student population. Using an iterative building-block approach, ISD then determines the training system design requirements for the weapon system.

Using ISD in the military services for weapon system training development is a slow, labor-intensive process requiring extensive manual data manipulation. For emerging weapon system designs, the Logistic Support Analysis Record (LSAR) is the primary source of design and logistics information. The LSAR database is governed according to Military Standard 1388-2A and is required for all weapon system acquisition programs. Logistic Support Analysis (LSA) is an iterative process that regularly updates the system's design and supportability information through all weapon system acquisition phases. However, neither the LSAR process nor the LSAR database is designed to easily provide information to the ISD process. As a result, using LSAR data to support ISD is difficult and time consuming.

Working Prototype

The Working Prototype design will permit flexible and efficient data manipulations to support the performance of timely and accurate ISD analyses. The prototype will include modules that provide system security, administration, utilities, and report generation, as well as ISD analysis. In the ISD analysis modules, the analysis will be documented on automated analysis worksheets. The ISD analyst will be supported through an effective user interface including decision support logic and meaningful presentations of LSAR and other analysis-related data. An audit trail will record ISD analysis decisions for subsequent review and modification.

The Joint Service Training Requirement DSS Working Prototype hardware and software requirements are detailed in this report. The Working Prototype will be an IBM PC-based system designed for easy use by ISD analysts. It will consist of single PC work stations, each operating independently, one PC per analyst. Capacity will be sufficient to store and manipulate LSAR and analysis data for one weapon system. Procedures to extract training data elements from the LSAR can be easily accomplished using commonly available equipment and minimal training. The LSAR data input to the Working Prototype will be provided in ASCII files on diskette or tape.

The working Prototype ISD analysis logic and database structure will be implemented in R:Base for DOS. R:Base is a user-oriented, fourth-generation language and database management system that combines an English-like, non-procedural language with a database that supports relational structures. A complete database facility using shared-relational files eliminates data redundancy while providing dynamic file manipulation and logical file views. R:Base offers the ability to easily adapt the working prototype programs to meet changing requirements, including alternate hardware implementations and additional ISD processes. This will facilitate modifying the prototype to accommodate additional requirements preceding full implementation of the Joint Service Training Requirements DSS. A one-time update to the Working Prototype is anticipated.

Benefits of the Working Prototype

The following are some of the benefits or improvements anticipated from the use of the Joint Service Training Requirements DSS Working Prototype, particularly on the front-end of weapon system acquisitions. The Working Prototype will:

- Allow better use of ISD analyst time through productivity improvements resulting from an automated LSAR data interface
- Provide better data control through standardized ISD analyses, data handling procedures, and security measures
- Reduce data handling errors through a standard LSAR data extraction routine
- Reduce the time required to progress through portions of an analysis by the automated transfer of interim analysis results from worksheet to worksheet
- Streamline ISD procedures
- Provide easier storage of ISD analysis results with an automated audit trail, thus minimizing the requirement for paper records and facilitating training system updates as the weapon system design changes
- Provide output report formats tailored to user needs, reflecting pertinent training information required for ISD coordination and decision making
- Provide automated output for direct inclusion in training equipment design specifications

Limitations of the Working Prototype

The Working Prototype is intended to demonstrate and test Joint Service Training Requirements DSS design concepts in an operational environment. It will provide an interim Joint Service automated ISD analysis tool and LSAR interface. While it will closely resemble portions of the final Joint Service

Training Requirement DSS implementation, it will contain only a subset of Joint Service capabilities. Some of the specific limitations of the Working Prototype are:

- Limited Flexibility. The Working Prototype only implements the ISD procedures of the 3306th Test and Evaluation Squadron (TES), Edwards AFB. The 3306th TES ISD process is an adaptation of the procedures documented in Air Force Manual 50-58, Instructional Systems Development, with particular focus on front-end ISD. The 3306th TES ISD process will greatly benefit from the automated interface with LSAR data. Additionally, although other task selection models exist, only the 3306th TES model will be included.
- Limited Range of ISD Functions. In order to automate the data handling, administrative, security, and report generation functions of the prototype, limited processes will be implemented in the preliminary Working Prototype. The prototype design will permit the easy addition of other ISD processes. A one-time update to the Working Prototype is planned to incorporate most or all of the remaining ISD processes.
- LSAR Change Data. No LSAR update capability has been implemented as a part of the preliminary Working Prototype. The ability to process LSAR change data will be provided as a more flexible LSAR-to-ISD data interface with the final Working Prototype update.
- Performance and User Interface. Specific performance requirements were not specified by the government for the Working Prototype. R:Base for DOS provides for rapid prototype implementation, with possible tradeoffs in operational efficiency and user interface. Performance requirements and an optimal user interface will be prime concern for the final Joint Service Training Requirements DSS design.

Working Prototype Documentation

Documentation for the Working Prototype will include a System Overview and a User Manual. The System Overview will include a description of the Working Prototype, directed toward management and staff personnel who have no need for detailed technical information concerning system implementation or operation. The Working Prototype User Manual will be directed toward supervisory and operator personnel who are primarily interested in detailed procedures of the computer program. The Joint Service Training Requirements DSS Data Dictionary will also be used as the data dictionary for the Working Prototype.

Training

Training for the Joint Service Training Requirement DSS Working Prototype will consist of initial training for ISD supervisors and analysts. The training will include basic Working Prototype data entry techniques, LSAR and other data file manipulations, user options available during prototype operation, and standard output generation. Training materials will be provided to personnel to be trained in advance of the scheduled classroom training. Training will include actual hands-on Working Prototype use.

Computer hardware/software required for training will be provided by the government at the training site. Training will be conducted as soon after prototype installation as practical.

Interfaces

The Joint Service Training Requirements DSS is a stand alone workstation. The only interface is with LSAR data received from manufacturers of new weapon systems. Applicable Training Data Elements are extracted from the manufacturer's LSAR data tape. The contents of the tape are design and logistics information for new weapon systems. The tape is prepared according to Military Standard 1388-2A. This tape is a required submittal by manufacturers of all new weapon systems. Regular updates, meaning additional tapes consisting of current and additional design and supportability information, are also required. The magnetic tape is read by a mainframe computer with a compatible tape drive. From there, the data is imported to an IBM-PC and mapped into the system's relational database.

The process that captures data from the LSAR tape must handle several thousand records that are expected to occur on the tape. An automatic determination of size, in terms of the Training Requirements DSS database, is necessary in advance to ensure that space is available on the database for the new data. Format, data codes, unit of measure, and range of values are prescribed in Military Standard 1388-2A.

Discussion

The Joint Service Training Requirements DSS will eliminate some of the inherent inefficiencies in the existing ISD procedures by automating many of the analysis and data handling steps and by automating the interface between the ISD process and LSAR data. The end result will produce a system that streamlines ISD functions, eliminates redundant tasks and data, improves analysis quality and effectiveness, and reduces cost and lead time. This Training Requirements DSS will provide the Department of Defense and industry with a standardized application of the ISD process and an LSAR data package tailored and standardized to support each Armed Service's training development.

References

- Department of the Army. (1975). Interservice procedures for instructional systems development. (Phases I through V). TRADOC PAM 350-30. Ft Benning, GA: U.S. Army Combat Arms Training Board and Florida State University, FL: Center of Educational Technology.
- Department of Defense. (1986). DoD requirements for logistic support analysis record. MIL-STD-1388-2A. Washington, DC.
- Department of Defense. (1985). Military Standard, MIL-STD-1388-2A (Navy). Washington, DC: Department of Defense.
- United States Air Force. (1987). 3306th procedural handbook.

USAF Integrated Manpower, Personnel and Comprehensive Training & Safety (IMPACTS) Program

Elaine Howell, Major, USAF
DCS Product Assurance and Acquisition Logistics (PL)
HQ Air Force Systems Command, Andrews AFB, MD 20334

Many contractor studies and internal inspections over the past several years have indicated to senior level planners that manpower, personnel and training (MPT) are high drivers in weapon system life-cycle cost, and that the design and acquisition process does not consider MPT concerns early or adequately enough for them to influence design trade-offs and guarantee supportability. Ever increasing interest, due to end-strength constraints, budget cuts, and Congressional requirements, has led the Air Force to implement the Integrated Manpower, Personnel and Comprehensive Training & Safety (IMPACTS) Program.

The 1987 Defense Authorization Act required major programs to submit projected manpower requirements to Congress 90 days prior to approval for Full Scale Development. Changes in Title 10, US Code, Section 2434, reduced this to 60 days. A draft Department of Defense Directive on Manpower, Personnel, Training and Safety (MPTS) in Weapon System Acquisition was issued in the summer of 1988. This directive specifies the content and format of the Manpower Estimate Report (MER) to be submitted to the Assistant Secretary of Defense, Force Management and Personnel (ASD ((FM&P))), and adds a requirement for an iterative MPTS Profile to be submitted at Milestone "0" and at every milestone thereafter.

In the summer of 1987, the Air Force Systems Command (AFSC) Commander, General Randolph, directed a colonel-level planning committee to study the problem and consider the possibility of an Air Force MPT Broad Area Review (BAR). After thorough consideration of history, and review of current trends and status, the committee decided not to recommend a BAR, but did recommend the implementation of IMPACTS. Four basic objectives were established:

1. Expand the model MPT organization formed at Aeronautical Systems Division (ASD), Wright-Patterson AFB, Ohio, to the other AFSC Product Divisions. The charter of the Directorate for Manpower, Personnel and Training (ALH) is to train MPT analysts, and matrix them into the system program offices (SPO) to assist the Program Managers in making trade-off decisions.
2. Increase emphasis on Manpower and Personnel, and Training and Training Support elements of Integrated Logistics Support (ILS).
3. Institutionalize MPTS in the systems engineering process. Make designers more aware of supportability requirements when considering design trade-offs, to ensure MPTS is an integral part of that consideration.

4. Provide senior-level, top-down direction and support for the program. The planning committee also recommended that the Principal Deputy, Assistant Secretary of the Air Force (Acquisition) (SAF/AQ) become the program office of primary responsibility.

As a result of these recommendations, the planning committee tasked ASD/ALH to more fully develop the IMPACTS concept and draft a Memorandum of Agreement to implement the program Air Force-wide. By DoD direction, the program must encompass not only manpower, personnel, training and safety, but also human engineering and environmental health hazards. The DoD definition of the program concept is that of the "...human dimension of the complete defense weapon system." In other words, the man/machine interface.

IMPACTS is based on a three-level structure, with central steering committee oversight. The three levels consist of an IMPACTS Working Group at the Air Staff, IMPACTS Planning Teams in each SPO, and IMPACTS Focal Points at each major command (MAJCOM). The basic structure of IMPACTS was formulated with several objectives in mind: to require no new administrative or managerial manpower requirements; to involve those offices which have had individual responsibilities for manpower or personnel or training or safety or human engineering, and bring them together in an integrated planning forum.

SAF/AQ accepted the transfer of ownership from HQ USAF/DP, and appointed the chief of the Advanced Programs Division (AQPT) to manage and direct the Air Force-wide program. The steering committee is made up of colonel-level representatives from Air Staff agencies (DP, LE, PR, XO, IG), AFSC agencies (PL, XT, SD, IG) and focal points from the MAJCOMs. The committee, chaired by a colonel from SAF/AQPT, will set IMPACTS policy for the Air Force, consistent with DoD and AF regulations and requirements.

SAF/AQPT also heads the Air Staff IMPACTS Working Group, which will act as the approval and review authority for all MPTS/IMPACTS documentation. This action officer working group receives and evaluates inputs from program documentation such as Statements of Need (SON), System Operational Requirements Documents (SORD), Mission Needs Statements (MNS), Program Management Directives (PMD), Test and Evaluation Master Plans (TEMP), and Integrated Logistic Support Plans (ILSP). They also review and approve the Manpower Estimate Report and the MPTS Profile required by DoDD 5000.xx, as well as the IMPACTS Program Plan (IPP).

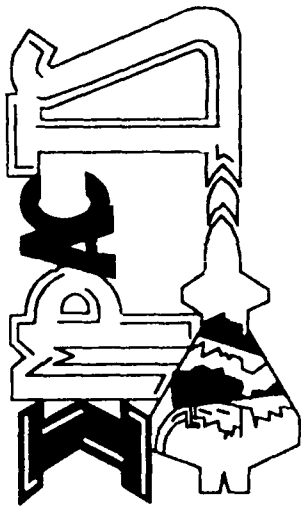
Existing Training Planning Teams at each SPO will be expanded into IMPACTS Planning Teams (IPT). M, P, S and Human Factors (HF) expertise will be added, and an MPTS analyst, matrixed from ALH, will act as integrator. The existing Training Development Plan will become the IMPACTS Program Plan, written by the IPT. The IPP will consist of stand alone plans for each of the program elements, an executive summary, and an integration plan. The plan is iterative and fluid; updated as changes and trade-offs are made in the program. The IPP becomes the single source document for MPTS information and documentation for each program. The team also writes the MPTS Profile. This team must establish the human resource implications of the tasks specified by the designers. From these, they must identify the "MPTS high-drivers": those tasks which require large human resource requirements, or generate potential safety problems or hazardous conditions. They make recommendations for

design trade-offs to change these factors. They eliminate them by redesigning the tasks, or by eliminating the human requirement to perform them. This includes changes in hardware, software, or procedures.

MAJCOM participation in the requirements and design process is considered vital. Each MAJCOM will appoint a focal point to sit on the colonel-level IMPACTS Steering Committee. This focal point will present an integrated MPTS viewpoint for the command, using inputs from the various players involved in determining system M & P & T & S requirements. Each MAJCOM will also appoint IMPACTS representatives to be members of the IMPACTS Planning Teams at each SPO. The grade of the representative will depend on the scope of the program. The focal points and representatives serve as information conduits to and from the steering committee and the working groups.

The DoD Directive allows each service component 120 days to publish an implementing regulation. ASD/ALH is drafting and coordinating an Air Force Regulation, that will be incorporated in the streamlined AFR 800-2, and will establish IMPACTS policy and requirements. The IMPACTS MOA has completed final coordination and is ready for signature. Changes to other existing AF regulations have been written and submitted. Policy letters, similar to those which established operating procedures for R & M 2000, are being written and coordinated. The feasibility and cost of the future expansion of the ALH concept to the other AFSC product divisions is under study, and a recommendation is expected by Spring '89. An IMPACTS newsletter will, hopefully, soon be expanded to a professionally published bulletin which will be distributed throughout the MPTS community. The AFSC Commander added an MPT chair to the Acquisition Strategy Panel to insure MPTS concerns are considered, and human factors and supportability are not traded for hardware without considerable justification.

The role of IMPACTS is two-fold: it is a program to integrate MPTS and HFE expertise to meet Congressional requirements, and it mandates a fully integrated relationship and a new spirit of cooperation between the equipment designers and the human factors experts. The IMPACTS short-term goal is to achieve the four basic objectives set by the planning committee without an additional manpower requirement. The IMPACTS long-term goal is to meet the AFSC Command goals: 1. Meet the user's need, 2. Maintain acquisition excellence, 3. Enhance technological superiority. IMPACTS will provide the user with more efficient and effective MPTS support for new and modified weapon systems.



INTEGRATED MANPOWER, PERSONNEL AND COMPREHENSIVE TRAINING & SAFETY

OVERSIGHT

DIRECTION

IMPACTS STEERING COMMITTEE

DODD

5000 XX

THREE LEVEL STRUCTURE

HQ USAF

WORKING

GROUP

SPO

PLANNING

TEAMS

MAJCOM

FOCAL

POINTS

IMPACTS

MOA

AFR 800-XX

ASD/ALH

MODEL MPT ORGANIZATION

POLICY
LETTERS

MATRIX

ANALYSTS

TRAIN

MPT'ERS

DEVELOP

TOOLS

MPT

RESEARCH

USAF AERONAUTICAL SYSTEMS DIVISION'S MODEL MANPOWER, PERSONNEL AND TRAINING ORGANIZATION-- AN UPDATE

Lt Col Frank C. Gentner
Chief, Analysis Division
Manpower, Personnel, and Training Directorate (ASD/ALH)
Acquisition Logistics Deputate
Aeronautical Systems Division (Air Force Systems Command)
Wright-Patterson AFB, Ohio

ABSTRACT

Based on the critical need to enhance consideration of manpower, personnel and training (MPT) factors early in the weapon system acquisition process (WSAP), the United States Air Force established a model MPT integration organization at the Aeronautical Systems Division (ASD) at Wright-Patterson AFB, Ohio. This organization was created by a memorandum of agreement between Air Staff (HQ USAF), Air Training Command (HQ ATC), and Air Force Systems Command (HQ AFSC). They dedicated 36 manpower positions to this purpose. The organization was chartered to study, recommend, and test ways that the Air Force's most expensive asset -- people -- can more fully affect weapon system design. This paper has four major sections: (I) MPT integration problems to be solved, (II) the new organization's mission and functions, (III) the Directorate's concept of operation, and (IV) current projects and activities the organization has undertaken to integrate MPT considerations into the WSAP.

SECTION I - MPT INTEGRATION PROBLEMS TO BE SOLVED

One third to one half of every dollar spent on Air Force (AF) weapon systems over their life cycle goes toward funding manpower, personnel and training. In today's environment of declining manpower authorizations, decreasing skilled labor pools, and severe budgetary constraints, MPT integration offers potential multi-billion dollar savings by more effectively using our people resources. Beginning in the early 1980s, a number of studies (Defense Science Board Study, Air Force contract studies, GAO reports, and AF Functional Management Inspections) served to bring MPT problems into clearer focus. These studies clearly demonstrated that early influence of design for MPT supportability was key to making substantive impacts of MPT requirements. Benefits derived from early MPT optimization include reduced life cycle cost, reduction of human constraints in weapon system performance, and enhanced combat capability.

In March 1986, an MPT Memorandum of Agreement (MOA) established the AF model MPT organization charged with resolving identified MPT integration problems. The AF WSAP needed a systematic approach to manage the MPT integration. MPT planning was generally fragmented and ill-timed, with MPT factors being considered too late to influence system design. Usually, the major efforts to analyze MPT impacts were in the full-scale development phase, after most life cycle costs are fixed. Planning efforts were not comprehensive enough. MPT goals were not clearly defined for contractors. Manpower estimates needed to be identified accurately at the outset, with full consideration given to the specialties to be involved. Data for MPT analysis was not available in a usable format to allow trade-offs, nor were they available through on-line data transfer networks to expedite data use. Existing MPT analysis tools were segmented and lacked capability to support management decisions. Training and training equipment were not adequately or consistently funded, developed, or procured concurrently with their weapon system. MPT management was highly decentralized, with no organization responsible for MPT integrating functions. Little effective direction was given to system program managers on MPT issues by Air Staff or the implementing commands. Lack of controls over the MPT process resulted in duplication of effort, higher cost for weapon systems, and ill-defined and late-to-need aircrew and maintenance training systems.

SECTION II MISSION AND FUNCTIONS OF THE MPT DIRECTORATE

The AF model organization is located in the Deputy for Acquisition Logistics, Aeronautical Systems Division (AFSC) at Wright-Patterson AFB, Ohio. The MPT Directorate's mission is dedicated to improving analysis and integration of MPT issues in the acquisition cycle to ensure optimal MPT supportability. Major functions include: (1) Develops plans, policies and analytical tools to quantify MPT impacts on and of developing weapon systems, placing special emphasis on front-end analysis. (2) Employs analysis techniques, policies and procedures to ensure consideration of alternative MPT utilization concepts and systems designs, encouraging necessary trade-offs to optimize cost and force effectiveness. (3) Maintains contact with Systems Program Offices (SPOs) and other ASD offices to assure participation in their studies and analyses. (4) Maintains liaison with research organizations to promote research into areas beneficial to MPT. (5) Functions as the ASD focal point in directing activities of MPT analysts assigned to SPOs to advise, assist and provide technical information and analysis support. (6) Provides

direction and leadership to obtain necessary information systems. (7) Advises the ASD Commander and HQ AFSC through the Deputy for Acquisition Logistics (ASD/AL) on MPT matters. (8) Advises the MPT Steering Committee on the Directorate's progress in meeting the organization's objectives. (9) Maintains liaison with key AF MPT organizations, such as AFMPC, AF Manpower Engineering Agency, ATC, USAFOMC, and other Steering Committee offices to exchange MPT information, data and concepts. (10) Serves as the "model" Air Force MPT integration organization. Ensures that a comprehensive management plan is developed and updated for use in developing an AF-wide MPT program. (11) Publicizes MPT integration issues and successes using available publication media including news releases, news letters, briefings, and publications.

SECTION III - MPT DIRECTORATE CONCEPT OF OPERATION

STRUCTURE. The MPT Directorate is structured to provide a central group of MPT personnel in the home office with visibility across program lines to ensure support and organizational objectives are attained; provide expertise for specific programs and/or issues during high demand periods (preparation and review of Requests for Proposal [RFPs], source selections, etc.); and provide matrixed MPT expertise to work the program-specific issues as the weapons system progresses through the WSAP. ALH's home office selects, trains and matrixes the MPT analysts to selected SPOs. ALH also provides continuing training, advice, analytic support, and counsel. The home office is responsible for career enhancement, and supporting them on the more complex MPT analyses. For those SPOs not having matrixed personnel, ALH provides staff assistance and MPT analysis on an as required basis. The home office also provides ASD policy updates and guidance, reviews Statements of Operational Need (SONs) and System Operational Requirements Documents (SORD), RFPs, Statements of Work (SOWs), and assists in developing an AF-wide MPT program for SAF/AQ. The Directorate addresses MPT supportability in both planned and present acquisitions.

MATRIXED PERSONNEL. Matrixed personnel work for the Director of Logistics (DOL) to ensure MPT issues are fully addressed. Assisting the DOLs with MPT expertise in working the Integrated Logistics Support (ILS) and Training Development Plans (TDP), they support the Chief System Engineer in identifying the full MPT ramifications of the various design issues faced by the SPO. As a specific weapon system transitions through the acquisition process, they aid the SPO in establishing an MPT baseline, and ensure all affected agencies are aware of the impacts upon that baseline as design, operational, or maintenance concept changes are proposed. They work interactively on all design issues having MPT implications, closely with SPO-matrixed engineering and logistics personnel. Existing MPT analysis tools are being used to compare, project, and assess different design options and operational and maintenance scenarios for their relative MPT impacts and life-cycle costs. Extensive use is being made of currently available methodologies and tools such as Logistics Support Analyses (LSA), LSA Records (LSAR) reports, the Logistics Composite Model (LCOM) and Cost Oriented Resources Estimating (CORE) model outputs, and other MPT data sources and models. The MPT Directorate develops contractual document statements which ensure that contractors are responsive to MPT concerns, and is developing standards for evaluating MPT proposals during source selection.

ASSISTING ASD ORGANIZATIONS. Several key ASD organizations have a special relationship with ALH. These include ASD's Deputy for Engineering (EN), Training Systems (YW), Development Planning (XR), the Safety Office (SE), and the ATC Operating Location (ASD/TTGT).

MAJCOM COORDINATION. In keeping with AF Systems Command policy, ALH works closely with the using MAJCOM to identify and satisfy user requirements during the WSAP. The using MAJCOM is required to consolidate manpower estimates prior to Milestones II and III to satisfy the statutory reporting requirements. Therefore, prime contractor, AF Logistics Command, SPO, and Air Training Command manpower estimates are consolidated by the using MAJCOM. They provide the manpower estimate report (MER) to HQ USAF, who forwards it through the Office of the Secretary of Defense to Congress. Close cooperation between ASD/ALH, EN, the SPO, and MAJCOMs is essential to ensure these important estimates are accurate and that the system design meets user needs within realistic MPT constraints. These key players must closely coordinate to achieve an optimum balance between manpower, system design, and mission effectiveness.

WSAP TIMELINE ACTIONS. Another way of describing ALH's concept of operations is to examine major MPT functions by acquisition phase

PRECONCEPT: Develop MPT constraints and goals, define problems for resolution with the new system, identify MPT analyses and trades to be examined, examine new technologies, participate in planning projects, and develop source selection criteria.

CONCEPT: Explore MPT alternatives, examine implications of design trades, develop alternate MPT concepts, influence design, and develop source selection criteria.

DEMONSTRATION / VALIDATION: Evaluate MPT implications of alternate systems, recommend changes, refine MPT concept and AF Specialty (AFS) structure, help estimate manpower, order data for training development, plan MPT tests, and develop source selection criteria.

FULL SCALE DEVELOPMENT: Evaluate system for MPT issues, refine, finalize and publicize MPT concept and AFS structure, ensure training system developed for concurrent fielding with system, help finalize manpower estimate, and evaluate MPT tests.

PRODUCTION / DEPLOYMENT: Review test results for MPT implications, evaluate engineering change proposals (ECPs), develop MPT lessons learned, and validate MPT concept

LOGISTICS READINESS AND SUPPORTABILITY: Determine whether MPT estimates were on target, review ECPs for MPT impact, and develop lessons learned.

MAJOR UPGRADE OR REPLACEMENT: Determine total MPT costs of system, identify problems and opportunities, determine possible MPT savings of new/upgraded system, and develop MPT breakpoints at which new/upgraded system becomes incrementally improved.

SECTION IV - CURRENT ALH PROJECTS AND ACTIVITIES

Current ASD/ALH projects and activities can be grouped into three categories: (1) Improvement of MPT policy and awareness; (2) Development of analysis tools, data bases, and procedures for using them, and (3) Direct acquisition program (SPO) support.

MPT POLICY AND AWARENESS

With guidance from the MPT Colonel-level Steering Committee, ALH has been working a number of MPT policy and awareness initiatives. Foremost in these efforts is the IMPACTS program development.

IMPACTS PROGRAM DEVELOPMENT. Integrated Manpower, Personnel, And Comprehensive Training/Safe / (IMPACTS) was established as the official Air Force MPT and Safety integration program by the MPT Steering Committee in April 1988. The goals of the program are to integrate human-centered disciplines of MPT, safety, and human engineering to support development of mission capable systems that can be safely operated, maintained, and supported in present and future operational environments at the lowest life cycle cost and with the people who will be available. IMPACTS focuses on setting MPTS goals and constraints at the outset, participating in the trade process, measuring the MPTS impacts or system design, and ensuring that the MPT pipeline is ready to support the system when fielded. ALH drafted an IMPACTS regulation for inclusion in AFR 800-2, Acquisition Program Management. The IMPACTS draft regulation is presently in coordination. The basic forum which brings those concerned with human issues together is the proposed IMPACTS Planning Team (IPT), which produces an IMPACTS Program Plan (IPP). The IPT is critical to the interdisciplinary approach needed to integrate MPTS. The IPT is comprised of representatives of the implementing, supporting, and using MAJCOMs, who address system MPTS and HFE issues. They furnish issues and content for the IPP which stores essential MPTS data about the developing system. The IPP documents the IMPACTS strategy; identifies goals, requirements, constraints; and issues for each IMPACTS element. It specifies design trade-offs and studies necessary to obtain IMPACTS objectives; lists the major MPTS tests and evaluations; and outlines necessary milestones to meet these objectives.

COORDINATION WITH SAFETY AND HUMAN FACTORS INITIATIVES. Because safety and human factors issues can have large MPT implications, close coordination with the ASD Human Factors Engineering (HFE) Branch and Safety Office is essential. Both human factors and safety data sources can help identify high driver MPT issues. ALH is working closely with these offices to develop data-based ways of identifying MPT issues of consequence. By supporting issues of common interest, the three offices have a better chance of obtaining funding support for necessary analyses and of gaining implementation of HFE, Safety, and MPT positions. The IMPACTS program plan and planning team will reinforce this relationship.

REGULATORY GUIDANCE. ALH reviews acquisition and MPT regulations to determine consistency and their impact on MPT integration. Many ALH suggestions have already been incorporated into AF acquisition regulations, and more MPT direction will soon be published. In addition, the Military Standards (MIL STDs) or Military Primes (MIL PRIMES) and Data Item Descriptors (DIDs) furnish guidance to contractors and writers of contracts. Revision and streamlining will greatly facilitate MPT integration. ALH was designated as the AF representative to the two-year DoD MPT DID Review and Consolidation study, working in conjunction with the DoD Human Factors Technical Advisory Group.

DEVELOPMENT OF CSNAS NETWORK. While conducting a thorough review of regulatory guidance, ALH, working cooperatively with the Acquisition Logistics Center (AALC), developed an MPT time-phased roadmap. This roadmap is presented on the Computer Supported Network Analysis System (CSNAS). This computerized "PERT diagram" provides SPOs a model MPT network which shows what should be accomplished, while allowing it to be tailored to their program. CSNAS is a government-owned project management tool which meets regulatory requirements to maintain a networking system.

SYSTEMS MPT COURSE. While attempts have been made, it is clear that a course which fully covers manpower, personnel and maintenance training (as well as operator training) needs to be developed. As funds become available, ALH plans to monitor the contract for a comprehensive MPT course(s). The

course will need to be tailored to meet the needs of MPT managers, logistics, human factors, and other AF and industry personnel. A detailed MPT analyst course would teach ALH, SPO matrixed, and other acquisition personnel to conduct weapon system MPT analysis. As a stop-gap measure, ASD/ALH has constructed, and presents as requested, a two-day MPT course, which highlights MPT acquisition issues.

UPDATE AFIT, AFALC, and ASD ACQUISITION COURSES Working with AFIT, AFALC, and ASD training personnel, ALH has updated the courses used to educate and train ASD personnel on the acquisition process. ASD/ALH prepared a two-hour MPT training module for inclusion in these various courses, and has received positive feedback on the module's implementation.

DEVELOPMENT AND USE OF ANALYSIS TOOLS, DATA BASES, AND PROCEDURES

While the AF led development of advanced analytic tools and data systems valuable for assisting MPT integration, it has not consistently applied available tools and data systems within a coherent framework. ALH is attempting to apply existing MPT analysis tools and data bases sufficiently early in acquisition to influence system design.

LCOM USED IN AN ITERATIVE PROCESS AND AS SOURCE OF MAINTENANCE MANPOWER ESTIMATES. One of the most valuable manpower analysis tools for use on aeronautical systems is the Logistics Composite Model (LCOM). LCOM is the AF-approved manpower model which has been validated to produce aircraft maintenance manpower assessments through simulation. For aircraft maintenance, this method is far superior to and more accurate than the other manpower estimate procedures which add task times, rather than model the interactive cueing effects. LCOM uses the Maintenance Data Collection (MDC) system to furnish the crew size, frequency and maintenance tasks times for each aircraft system. Before inputting these data into LCOM, the MDC data are operationally audited by manpower personnel for consistency and accuracy. MDC data offers hard maintenance data on predecessor systems, which can be used in LCOM comparable system simulations of future aircraft maintenance. LCOM is used as a tool to identify MPT high drivers and aircraft maintenance manpower assessments throughout the WSAP. LCOM will be conducted at increasing levels of specificity from Preconcept sensitivity analyses on the predecessor system, to baseline system comparisons during Concept Exploration, to LCOM models based on LSA comparability input data during the Demonstration/Validation Phase and more final LSA/engineering data during Full Scale Development. LCOM is the official source of aircraft maintenance manpower estimates. In line with this concept, LCOM is being applied to the Advanced Tactical Fighter (ATF). The manpower implications of alternate AF Specialty structures (differing task assignments to career fields) are being examined. The consideration of alternate AF specialty structures is, itself, a result of initiatives by an ALH SPO matrixed analyst.

PERSONNEL AND OCCUPATIONAL SURVEY DATA. The Air Force has one of the most sophisticated personnel data systems available in any Service. It can describe AF military personnel in great detail by career ladder and has great flexibility for data manipulation to answer MPT questions with the Advanced Personnel Data System (APDS). Comprehensive Occupational Data Analysis Programs (CODAP) enables the USAF Occupational Measurement Center to describe the tasks and related task data about each AF career ladder. Both APDS and CODAP data bases feed the Occupational Research Data Bank (ORDB) which makes computer "runs" available through "on-line" modem access. The one drawback to these data is that they are presented by Air Force Specialty Code (AFSC), which may or may not be directly related to an aircraft system. Research is under way to alleviate this disconnect. In addition, as the Rivet Workforce AFSCs are implemented, most maintenance AFSCs will be tied to a weapon system through the 14-year point in a maintainer's career. As these new AFSCs are surveyed, CODAP data will become more weapon system specific. Once these changes and the research are complete, the rich AF personnel data resources can be made available for the personnel/AFSC analysis needed to input to the LCOM manpower analysis. Until that time, ALH plans to use these data as best possible in a manual mode.

LSA/LSAR ORDERED ITERATIVELY AND SOURCE FOR TRAINING DEVELOPMENT. LSA/LSAR task data will be ordered iteratively from the prime contractors throughout the WSAP, beginning with Milestone 0 to provide data for addressing MPT issues. It is critical that the users tailor what they need, specify when each delivery is to take place, and that the medium specified is the most efficient/effective for immediate use. Later in the WSAP, complete and accurate LSA/LSAR reports are needed to address training development issues. If it is late, the training development system cannot ensure adequately trained personnel for the Initial Operational Capability (IOC). In the past, LSA/LSAR has received low priority. ALH will attempt to give higher visibility to LSA through the IMPACTS program plan so that the AF buys a total weapon system, including everything necessary for trained personnel to operate, maintain, and support it when fielded. To assist matrixed analysts in being aware of the most critical MPT LSA/LSAR to order, ALH is developing a handbook to describe when in the WSAP to order the most critical LSA/LSAR. In later phases of this project, ALH will determine MPT deficiencies in LSA/LSAR, as well as determine whether the government data given to contractors, the analytic methods, and the method of verifying LSAR are adequate. This long term project will feature interim updates to the MPT LSA handbook.

HUMAN FACTORING OF MAINTENANCE TASKS. In the past, most AF human factors efforts centered around the cockpit and crew positions. Because of the expense and lessened availability of maintenance manpower, equal emphasis is being given to human factoring of maintenance. The result will be maintenance tasks and equipment which reduce manpower requirements per unit, and increase the

efficiency of maintenance turn-around and wartime readiness of new systems, along with decreased probability of accidents or incorrect maintenance action. Validated human factor tools, which have been applied for years to cockpit design, could be applied more fully to maintenance. And both safety and human factor data bases could be used to focus attention on MPT high drivers in time to affect design. The IMPACTS program plan and working group will highlight the need for early application of human engineering.

AUTOMATED DESIGN TOOLS. Most prime contractors are now using Computer Aided Design (CAD) systems, and some have a human factor design tool like the AAMRL/AFHRL-developed programs, Crew Chief and COMBIMAN. These tools can impact on contractor engineers while they are in the design process. These automated design aids use a three dimensional manikin on the CAD screen, allowing engineers to "see" accessibility problems. By making these systems available to all prime contractors and requiring their use, many of the access problems experienced in the past can be overcome during the design process. ALH is encouraging their use and standardization into the design process. In addition, when the Defense Advisory Committee On Women's Issues In The Services (DACOWITS) raised a question on how well a developing aircraft would be at accommodating women, ALH used these tools in conjunction with ASD/EN, AFHRL/LR and AAMRL/HEG to provide a response before manual mockups were constructed.

MPTS ANALYSIS TECHNOLOGY CONTRACT. Realizing the need for data-based MPT decisions, the Human Systems Division (HSD), with the assistance of ALH developed and commissioned a major study to develop specifications for a current technology MPTS analysis system. The study will depict MPTS decisions in the acquisition process using Integrated Computer Aided Manufacturing DEFINITION (IDEF) discipline that shows decision/action inputs, constraints, resources needed, and outputs. MPTS decisions/actions will be linked to the best available MPTS analysis tool and data base. After specifications for a current technology analysis system are developed, a deficiency analysis will point the way to future HSD research needs for MPTS integration technology. This 16-month effort began in September 1988 at the kick-off meeting hosted by ALH. ALH will play a key role in furnishing information to the contractors and evaluating contract deliverables. After the specifications are delivered, ALH plans to apply the current-technology MPT analysis system to ASD programs.

AF CROSSWALK / FOOTPRINT. At the request of ALH, the Defense Training and Performance Data Center (TPDC) agreed to prototype the MPT "Footprint" of ASD/XR's Preconcept Long Term Planning Project, the Advanced Tactical Transport (ATT) which will replace some C-130s in the year 2000. Footprint is a series of data listings which identify the total MPT resources (or profiles) of a predecessor weapon system. TPDC is working with the USAF Occupational Measurement Center to construct an AF Crosswalk system which will allow automated look-up tables which link a weapon system with all the MPT resources that support it. When operational, the Crosswalk will support automated Footprints of predecessor systems for goal setting, MPT profiles, and comparability analysis. ALH is playing a key role in helping identify the essential data needed during each acquisition phase. This will ultimately have a sizable impact on the form the AF Crosswalk/Footprint takes.

DIRECT SUPPORT TO SPOs

The ALH matrixed analysts are influencing MPT-related design decisions in a variety of SPOs. Already, significant MPT supportability issues have arisen. ALH personnel played a major role in preparing programs to develop MPT systems to support the Rivet Workforce AF specialties, and prepare for the more consolidated Rivet Workforce of the year 2000 and beyond. They have raised issues of how consolidated specialties can increase manpower utilization rates while contributing to war-fighting capability and improved people/machine manpower ratios. ALH personnel have also helped SPOs and using MAJCOMs prepare for the Congressionally-required Manpower Estimate Report (MER), and are preparing a computerized format for use by all SPOs. Human factors issues which impact on personnel selection and classification have also been surfaced. ALH personnel have ensured that the training planning necessary for fielding complete training systems with the weapon systems was accomplished. The dedicated advocacy that ALH furnishes to SPOs enables MPT to be a player in the program office decision-making process.

SUMMARY

The Manpower Personnel and Training Directorate was established to test whether a long standing need to fully integrate MPT considerations into the acquisition process could be institutionalized by a model organization at the AF SC product division level. The initial cadre of 12 military personnel developed an MPT Plan, which is now being implemented. The goal of the organization's concept of operation is to demonstrate how existing MPT analysis tools, data sources and procedures can more fully improve consideration of MPT factors in the WSAP. The MPT Directorate is testing use of ECOM, CODAP, HFL, CAD tools, and LSA/LSAR. Also, ALH is encouraging development of needed MPT analysis tools and data bases by identifying voids and inadequate tools. In addition, it has reviewed acquisition documents

for inclusion of MPT requirements, and is using CSNAS as a management prototype to encourage timely requesting of MPT analyses and data. Regulation and MIL STD changes coupled with a three-tiered Systems MPT course will help encourage acquisition managers to consider MPT. And finally, ALH is developing the IMPACTS program to ensure MPTS are given appropriate planning and visibility. With this program, MPT goals and constraints will be developed, trade-off analyses conducted, and an optimal MPT support system will be developed under the influence of ASD/ALH and their MPT matrixed analysts. The IMPACTS Planning Team and IPP will highlight the important MPT issues within and outside the SPO.

REFERENCES

- Aeronautical Systems Division / Manpower, Personnel, and Training Directorate (1987, August). ASD/ALH Implementation Plan. Wright-Patterson AFB, OH: Author.
- Air Force Systems Command (AFSC/PLL) (1987, September). AFSC Commander's Policy (Network Analysis-based Scheduling System). Andrews AFB, MD: HQ AFSC/PLL.
- Akman Associates, Inc. (1983, April). Enhancing Manpower, Personnel, and Training Planning in the USAF Acquisition Process, Final Report. Silver Spring, MD: Author (AD-F630514).
- Defense Science Board (1982, December). Report of the Summer Study Panel on Training and Training Technology. Washington, D.C.: Office of the Undersecretary for Research and Engineering.
- General Accounting Office (1981, January). Effectiveness of U.S. Forces Can Be Increased through Improved Systems Design (Rep PSAD-81-17). Washington, D.C.: Author (AD A114237).
- General Accounting Office (1985, September). The Army Can Better Integrate Manpower, Personnel, and Training into the Weapon Systems Acquisition Process. GAO/NSIAD-85-154. Washington, D.C.: Author.
- Gentner, Frank C. (1987, October). Importance of Weapon System-Specific Occupational Survey Data. Proceedings of the 29th Military Testing Association Meeting. Ottawa, Ontario, Canada: Director of Military Occupational Structures, National Defence Headquarters.
- Stephenson, Robert W., & Gentner, Frank C. (1987, October). Manpower, Personnel, Training, and Safety Guidance and Control for Weapon System Acquisition (AFHRL-TR-87-31). Brooks AFB, TX: AF Human Resources Laboratory.
- United States House of Representatives, 99th Congress (1986, October, pp 165-166.) National Defense Authorization Act for FY 1987, Conference Report. Washington, D.C.: Government Printing Office (HR 99-1001).

ABOUT THE AUTHOR

Lt Col Gentner is currently Chief, Analysis Division, of the Manpower, Personnel, and Training Directorate (ASD/ALH), Deputy of Acquisition Logistics, Aeronautical Systems Division (AFSC), Wright-Patterson AFB, OH. He is responsible for developing the analytic staff and techniques to be used by the Directorate and coordinating research needs with the Laboratories to facilitate integration of MPT issues into the weapon system acquisition process. He holds a Bachelors in Psychology and Masters in Rehabilitation Counseling from the University of Florida. In addition, he has taken post-masters Industrial Psychology courses at St Mary's University. He has served as a personnel officer, an Air Training Command Technical Training instructor supervisor and course chief, Director of Training Evaluation for Defense Equal Opportunity Institute, occupational analyst, staffer, and planner at the USAF Occupational Measurement Center, and initial Chief of Plans for Training Development Service.

ACKNOWLEDGMENTS

The initial cadre of the MPT Directorate, working closely with Aeronautical System Division's Engineering, Training Systems, and Development Plans, as well as the Air Training Command Operating Location and the AF Human Resources Laboratory personnel made development of the MPT Directorate's plan possible in a relatively short period. I would like to acknowledge the MPT Directorate's talented cadre for its work in developing the plan under the guidance of its Director, Colonel Christopher A. Somers. In particular, I would like to acknowledge the contributions of Lt Col Paul Cunningham, whose insight into the acquisition process and knowledge of LCOM have been invaluable to the Directorate's staff. This paper is largely based on that plan, which is now being implemented. The Directorate's procedures will continue to change, as needed, based on the feedback from the home office staff, the MPT SPO matrixed personnel, assisting ASD officers, and guidance of the MPT Steering Committee.

Establishing a Relationship Between Training Resource Expenditures and Unit Performance

Brian J. Bush
U.S. Army Research Institute Field Unit
Presidio of Monterey, California

In FY 88 the Army has dealt with budget constraints directly impacting on training more than anytime in the past decade. As a consequence there is both an increased need and effort by the Army to objectively quantify their training dollar requirements and to specify impact of budget shortfalls on unit training readiness.

The Army currently uses a variety of systems to manage their training resources which, in turn, provide input for the command operating budget. For most budget submissions by units and installations, dollar requirements are based upon historical training resource expenditures with an inflation factor. Additionally, budgets of battalions averaged across 'like' battalions with exceptions made for major training events such as a refresher.

A problem with this method of budget development becomes apparent when the dollars are translated into unit training readiness. Instead of projecting the actual cost of training to meet a prescribed training readiness level, the training is adjusted to meet the available dollars. Therefore units may do less training from one year to the next, while their reported training readiness remains the same. As long as budget reductions are seen as having little impact on reported training readiness, there may be little reason to be surprised at further reductions in the budget for training.

This paper describes a research effort by the Army which will identify the relationship between training resource expenditures and unit performance which, in turn, is a part of unit readiness.

Objective

The objective of this study is to describe the relationship between training resources expended by units at their home station and their performance at the Combat Training Centers, with primary emphasis at the National Training Center (NTC).

Method

Variables

Current data collection efforts are focused on acquiring training resource expenditures at home station and the NTC. These resources, or predictor variables, include data on miles driven, ammunition fired, repair parts consumed, annual training days, use of training aids (UCOFT, ART2ASS, and MILES), training events conducted, and personnel turnover. Also included as a predictor variable is a comparison between terrain at home station and at the NTC.

The collection of criterion data will be based on the fidelity of available data from the NTC. ARI has had a major research effort ongoing at the NTC for several years. Other symposia at MTA are presenting further information about that research effort.

Timelines

The data collection plan is linked to a unit's rotation date at the NTC. Data has been requested by month, and by battalion for the six months prior to their rotation. Expenditures are also being collected for the time during rotation for the BlueFor and OPFOR. Data on personnel turnover is being collected up to fifteen months after a unit's rotation.

The availability of data may require the use of roll-up data by longer time intervals in lieu of the requested monthly intervals.

Design

The primary focus of analysis is the conduct of correlational analyses to determine the relationships between resources expended and unit performance.

Instrumentation

The only instrument constructed for the data collection is a questionnaire designed to rate the similarity of terrain between home station and NTC.

The type of data collected from this instrument is similar to the more detailed information being collected by a "Determinants" contract effort for ARI on personnel characteristics, leadership, cohesion, and training.

Discussion

This research effort is a significant step toward a quantified and objective description of the training required for units to perform their mission at estimated levels of competency. From this step and subsequent refinements of data collection and analysis, the Army chain of command will be increasingly confident in the accuracy of budget submissions and better able to identify the impact of budget constraints and shortfalls.

Training Systems Analyst: The Changing Role
of the Behavioral Psychologist¹

Major Conrad G. Bills
Lt Nancy J. Fakult
Lt Z. Nagin Ahmed
James E. Brown
Aeronautical Systems Division (AFSC)

Abstract

The expansion of the Air Force into contracted aircrew training systems has created a void in the operational and engineering communities which is being filled by behavioral psychologists. This role changes the traditional human factors orientation from man-machine interface into a role of systems integration. New areas for behavioral psychologist preparation include the integration of the instructional systems development process with systems engineering and logistic support analysis. This gives a total training systems perspective. This perspective must include an understanding of the operational community in which the weapon system will be fielded as well as the acquisition community through which the attendant training system will be procured.

The Air Force has been moving rapidly in the direction of contracted aircrew training systems. During the past few years the Military Airlift Command (MAC) has initiated contracts to industry for modernizing their existing aircrew training systems. Psychologists out of the Air Force Human Resources Laboratory (AFHRL) were brought in as consultants to apply the results of their training effectiveness research for the C-130 Weapon System Trainer (WST). Nullmeyer & Rockway (1984) reported that in order to take advantage of the advanced training potential of the WST, simulator training must be integrated into the overall training system. Since this required a review of the entire training system, MAC chose to contract through AFHRL. The contractor, Seville Training Systems, was tasked to develop a model aircrew training system concept. Seville psychologists looked at the full continuum of aircrew training in a totally integrated system. Highburne, Spears, & Williams (1986) reported a system which would provide training program development, training delivery, training management, evaluation and training analysis, and training support. This system was to serve as a prototype in MAC and SAC aircrew training for large body aircraft.

1

The opinions and views of the authors do not represent official policies of the Department of Defense.

For new weapon system procurement, the Aeronautical Systems Division (ASD) has contract efforts concurrent with the weapon system contract to develop the attendant aircrew training systems (ATS). Two of these are the Tactical Air Command (TAC) advanced tactical fighter (ATF) and the MAC C-17 ATSS. For each program, behavioral psychologists are advising on the instructional systems development integration with systems engineering and logistics support analysis. This expanded role of the behavioral psychologist has brought the total training system perspective to weapon system acquisition.

Discussion

The speed at which the contracting of aircrew training systems entered the operational and engineering communities created a void that they were not prepared to fill.

Operations was concerned with the attendant problems in the present training system, but their limited and declining financial/personnel resources demanded commitment to their primary missions. Their perspective was, therefore, limited to their operational experiences. Their understanding was in the context of how they were taught and how they had learned to train others. This perspective was primarily within a wing with little vision of the full continuum for their training system. They were generally unfamiliar with state-of-the-art training technology and methodology and the associated terminology. Novel approaches, though of interest, were dismissed as unrealistic in the present environment.

Engineering was focused on the hardware and supporting software to derive the training devices for the weapon system. They depended upon operations to provide the training system for which the weapon system and the training devices would be integrated. Lead time for training device acquisition, particularly for full mission simulation, was increasing. Training requirements analysis became more than an assessment of what would be trained in the training device (Bills & Nullmeyer, 1985; Bills, 1987). The scope exceeded their experience. Instructional systems development process was imposed upon them without any correlation with their traditional systems engineering process (Fakult, Pfledderer, & Bills, 1988). New terminology such as formative and summative evaluation was foreign. They perceived instructional technology, particularly in the area of courseware, as not something they could approach with the rudiments of engineering.

From the Department of Defense also came the impetus to drive manpower, personnel, and training (MPT) issues earlier in the acquisition process. Even as early as concept development, questions are to be asked such as what will be the manpower requirements expected of this new weapon system, what will be the new distribution of personnel, and what will be their new training requirements. The increasing scope of the problems which must be dealt with by the acquisition community has caused them to reach out for additional resources to meet the demand. Behavioral psychology has been a resource called upon to help fill the void.

Behavioral psychologists have been trained in the fundamentals of behavioral observation, a basic tenant of educational technology. They also understand principles of behavioral analysis and research which are key to assessment of a contractor's training system development.

Behavioral psychologists have background in learning theory which is the basis for training system modeling. They are attuned to training technology, task and occupational analysis, and the environment in which a person is to achieve desired behavioral outcomes. Behavioral psychologists have the perspective to bridge the gap between the operational and engineering communities and the training system contractors.

Even though the behavioral psychologist is best suited to meet the new demand, he/she also has a training need. The human factors orientation of man-machine interface needs to be expanded into a systems engineering orientation. The behavioral psychologist needs an understanding of the Air Force instructional systems development (ISD) process and how the ISD process relates to the systems engineering process. Additionally, he/she needs to understand the integration of these two processes with logistic support analysis (LSA). With this understanding, the behavioral psychologist can then gain the important total training system perspective.

Finally, the behavioral psychologist needs an understanding of both the operational community in which the training system is to be fielded and the acquisition community procuring the system. This understanding is important since concurrency is the current policy. The acquisition strategy is to procure the training system concurrent with the weapon system. The training system, including full mission simulation, is to be in place ready to go when the weapon system is delivered. In order to make this happen, the behavioral psychologist is using effective observation skills to focus on the training need and then assist in translating this need into the integrated ISD/systems engineering process.

The industrial component of the acquisition community has centers of behavioral psychological expertise in training system development. Given the understanding of the operational need, industry has demonstrated their capability to produce the training system. The expanded role of the Air Force behavioral psychologist facilitates the interface of this industrial component with operations and engineering in the training system acquisition process.

As we look to the future, how can the preparation programs for behavioral psychologists better develop them to meet the expectations of this expanded role in training system acquisition?

Conclusion

The following six elements should be considered in the preparation of behavioral psychologists for their expanded role in training systems:

First, the human factors preparation should also include an introduction to the systems engineering process which relates this process to its derivative ISD process.

Second, the behavioral psychologist being assigned to either the operational or acquisition communities should attend an ISD course or criterion referenced instruction (CRI) course en route to their assignment.

Third, he/she should visit at least three operational settings. These operational settings should include wing operations and training as well as

initial qualification/combat crew training school (CCTS) with full mission simulation (WST) facilities.

Fourth, upon assignment he/she should visit a few contractor operated aircrew training systems (ATS) such as SimuFlite, the Center for Advanced Airmanship, the C-5 ATS or the C-130 ATS.

Fifth, the behavioral psychologist assigned to the acquisition community should immediately complete the course on introduction to acquisition management (SYS 100 or equivalent).

Sixth, during the assignment the behavioral psychologist should regularly attend the Interservice/Industry Training Systems Conference (I/ITSC) and periodically attend other related conferences to keep attuned to current developments and maintain contacts across the training systems arena. He/she should also be a member of the Training System Design Subcommittee of the Training Effectiveness Working Group (AFHRL & ASD).

References

- Bills, C. G. & Nullmeyer, R. T. (1985). Application of model aircrew training system (MATS) to B-52 combat crew training. Proceedings of the Military Testing Association Conference. San Diego, California.
- Bills, C. G. (1987). Integration of weapon system trainer (WST) for large body aircraft into combat crew training. Proceedings of the National Aerospace Electronics Conference (pp. 1054-1059). Dayton, Ohio.
- Fakult, N., Pfledderer, J. A., & Bills, C. G. (1988). Comparison of instructional system development (ISD) with systems engineering from the training system perspective. Proceedings of the National Aerospace Electronics Conference (pp. 1541-1543). Dayton, Ohio.
- Fishburne, R. P., Jr., Spears, W. D., & Williams, K. R. Design specification development for the C 130 model aircrew training system: final report (AFHRL-TR-86-51). Williams AFB, AZ: AFHRL Operations Training Division.
- Nullmeyer, R. T. & Rockway, M. R. (1984). Effectiveness of the C-130 weapon system trainer for tactical aircrew training. Proceedings of the 6th Interservice/Industry Training Equipment Conference (pp. 431-440). Washington, D.C.

A STUDY OF BEHAVIOR MODELING IN MANAGEMENT TRAINING

Dr. Phyllis Peters Marson
Federal Aviation Administration
Center for Management Development

Introduction

The study developed out of a need for research into the effect of behavior modeling training on job performance in organizational settings. The objective of this research was to investigate the relationship between behavior modeling as a method of training for problem solving and the on-the-job performance of managerial trainees.

The participants in the study were 150 Federal Aviation Administration (FAA) employees attending the Supervisor's Course, Phase I (SCI) at the FAA's management training school. The initial supervisor's course is a 100 hour pass/fail resident training course mandatory for all newly-selected agency supervisors. Selectees must successfully complete SCI within the first year of appointment. The trainees represented every regional and organizational function within the FAA and ranged in age from 27 to 59, with a mean age of 41. Fifteen of the participants were females, 135 males.

Typically, the trainees had begun their careers as technical specialists. Their exposure to managerial theories and techniques prior to attending SCI was limited. A Job Function Analysis and a Job Competency Analysis had been completed on the duties of first-line supervisor in the FAA to determine the skills, knowledges, and abilities necessary for management positions. Skill in mutual problem solving emerged as one of the essential competencies that differentiate between effective and ineffective managers.

A 12-hour competency based unit on mutual problem solving was developed for the SCI course using behavior modeling technology. The research was conducted on this component of the training process.

Instrumentation

A 27 item self-report instrument was developed to assess the trainees' interpersonal behaviors when they were engaged in mutual problem solving activities on the job.

The instructional objectives, subject matter content, and end-of-unit knowledge test were analyzed for internal consistency and adequacy of coverage within the problem solving unit. Each item on the instrument reflected a behavior indicator drawn from the generic four-stage mutual problem solving model presented in the

modeling display in the instructional unit. This problem solving model included: (1) an opening for establishing rapport and making a tentative statement of the problem, (2) a problem agreement stage in which there is mutual agreement upon the problem and its cause, (3) an exploration of alternative solutions, and (4) the conclusion stage in which a solution is selected and follow-up accountability accepted. The response scale used on the instrument was a six point forced-choice scale with number and word descriptors.

Collection of Data

Before beginning training, and again 10 weeks after training, the participants were asked to complete the instrument on problem solving behaviors. Employees from the trainee's work unit were asked to complete a subordinate assessment on the trainee post-training only. Anonymity was guaranteed the employees by requiring only the name of the management trainee. All instruments, except the initial pre-course instrument, were mailed with self-addressed envelopes enclosed for ease of return. Feedback on the participant's pre-and post-course assessment and on their subordinate's assessment was promised to the trainee if there were a minimum of three responses from the participant's work unit.

The instructional unit ended with a 3-hour period of individual evaluations and feedback, and an objective end-of-unit, paper and pencil, multiple-choice test. The scores from this test were collected and added to the pre- and post-course data.

A control group consisting of 25 Supervisory Identification and Development Program (SIDP) employees completed the problem solving instrument initially, and again 10 weeks later. The employees in this group were part of an FAA initiative designed to change the way the agency identifies, selects, and develops candidates for supervisory positions in air traffic. The control group had nominated themselves as potential supervisors and gained concurrence of five peers in addition to their immediate supervisor. They had gone through an assessment process that resulted in being rated as "ready-now" for supervisory selection. None in the group had been a supervisor, nor had any attended the SCI course. The data from this group were collected through the mail. The reliability of the questionnaire for the control group equaled .94 pre-course, and .97 post-course.

Data Analysis

Analysis of the data for this study was accomplished by comparing the means by analysis of variance (ANOVA), t-tests, analysis of co-variance (ANCOVA) where necessary to statistically control for pre-course differences, and the Pearson Correlation Coefficient.

Conclusions

The analysis of the data on the effect of behavior modeling training on job performance supports the following conclusions.

First, findings from the study indicated that the participants transferred to the job the mutual problem solving behaviors modeled in the classroom (Table I). This finding supports earlier research on applications of the behavior modeling approach to training.

TABLE I

Summary Table of Means, Standard Deviations,
and Percentage of Return

<u>Problem Solving Instrument,</u> <u>Pre-Course</u>		<u>Percentage</u> <u>of Return</u>
Participant	Mean = 110.658 SD = 13.883 n = 73	
Control	Mean = 121.952 SD = 14.379 n = 21	
<u>Problem Solving Instrument,</u> <u>Post-Course</u>		
Participant	Mean = 123.306 SD = 13.363 n = 72	48.0%
Su' rdinates	Mean = 120.185 SD = 17.134 n = 73	48.6%
Control	Mean = 123.500 SD = 15.954	67.0%

The second conclusion concerned the findings related to the control group. The participants showed significant gains in mean scores on the self-report questionnaire pre- and post-training, the control group demonstrated only negligible gains. The study demonstrated that the treatment made a difference; however, the small return of the control group (n=14, 68 percent return rate) post-course makes it difficult to be confident about the

effectiveness of the control in this study. That group also assessed their skills at a higher level pre-course, demonstrating very little change in mean scores in the ten weeks between responses. Even with the small n, however, this conclusion is in agreement with that of Latham and Saari (1979) who found performance change in supervisors trained with behavior modeling, but no change in the control group until they too received the training.

The third conclusion concerned the findings relevant to the effectiveness of self-report as a method of evaluation. The study indicated that there was no difference in the self-assessment of the managerial trainees and the subordinate assessment by employees with whom the trainees had problem solved on the job. This agrees with previous findings that: "...individuals possess an extensive data base from which to draw inferences about themselves, a much larger base than even the most ambitious external evaluator is likely to develop" (Shrauger & Osberg, 1981, p. 322).

Further findings in the study led to the conclusion that there was no difference in self-reported mean scores pre-training of supervisors who had been in their positions various lengths of time (from 0 months to twelve months). See Table II. All participants were within the first year of selection. On the basis of a small number of participants who had been supervisors six to nine months before training, however, there was an indication that those with more experience initially, showed more performance gain post-training on the job. There are significant training implications for the timing of supervisory training should these findings be supported by further research.

TABLE II

ANOVA of Means of Participants With Varying
Lengths of Time as Supervisor Pre-and Post Course

	n	Pre	n	Post
No previous experience	17	112.353	17	122.176
0 - 6 months	47	109.766	46	121.630
7 - 12 months	9	112.111	9	134.000
		F = .410		F=2.34

Finally, the study indicated no correlation between performance on a paper and pencil end of unit knowledge test and performance on the job. The lack of variance on the test (mean = 96.8, SD= 5.49), however, makes it untenable to conclude that there is no relationship. Further research is necessary to determine what link, if any, this variable has to job performance.

Recommendations

This present study needs to be replicated with a time period of at least six months post-training to allow the trainees more opportunities for applying the problem solving behaviors modeled in the classroom. Many of the 150 trainees who completed questionnaires pre-training, but not post-training, reported that ten weeks was not long enough for them to engage in problem solving activities with a sufficient number of employees. A longer time period for the study should give all members in a work unit the opportunity to participate in the assessment of the trainees' job behaviors. The long term effects of the training as well as the relatively short term effects need to be researched as well as the operational and economic impact of using self-report and subordinate assessment in performance appraisal.

Bibliography

- Latham, Gary P., & Saari, Lise M. (1979). Application of social-learning theory to training supervisors through behaviorial modeling. Journal of Applied Psychology, Vol. 64, No. 3, 239-246.
- Shrauger, Sidney J., & Osberg, Timothy M. (1981). The relative accuracy of self-predictions and judgments by others in psychological assessment. Psychological Bulletin, 90, 322-351.

Modeling the Costs and Benefits of Alternative Training Interventions

Michael D. Mumford
Georgia Institute of Technology

Joseph L. Weeks
Air Force Human Resources Laboratory

Francis D. Harding
Alexandria, Virginia

Edwin A. Fleishman
George Mason University

Because training is held to play a crucial role in ensuring adequate job performance, military organizations have invested billions of dollars in the design and implementation of training programs (Goldstein, 1986). This investment, and the foundation that training provides for effective performance, underscores the need for systematic evaluation efforts intended to establish the effectiveness of alternative training interventions (Campbell, Dunnette, Lawler, & Weick, 1970). Although a complex field in its own right (Goldstein, 1986), program evaluation and test validation share a common concern in that they both seek evidence indicating that the interventions under consideration lead to enhanced performance (Guion, 1965; Kirkpatrick, 1959). For some time, however, it has been recognized that the magnitude of the performance gains brought about by personnel interventions does not provide a fully adequate index for appraising organizational utility due to factors such as intervention cost, the relative value of performance gains, and the number of individuals exposed to an intervention over time (Brogden, 1946, 1949; Cronbach & Gleser, 1965; Schmidt, Hunter, McKenzie, & Muldrow, 1979).

Few studies have examined the utility of alternative training interventions (Goldstein, 1986; Schmidt, Hunter, & Pearlman, 1982). Four considerations, however, argue for the extension of traditional program evaluation efforts to incorporate utility considerations. First, like selection tests, even training interventions yielding relatively small improvements in aggregate performance may have great value given substantial variability in the value of performance differences, low cost, and applicability to a number of individuals over a substantial period of time. Second, utility models provide an attractive vehicle for integrating the multiple, often qualitatively different, kinds of criterion performances that must be considered in many program evaluation efforts (Goldstein, 1986; Kirkpatrick, 1959). Third, trainers are often confronted with the need to choose among a variety of alternative interventions and so evidence bearing on their relative value might prove useful in instructional systems design (Cronshaw & Alexander, 1985). Fourth, utility data might do much to further the development of general principles concerning the kinds of interventions most likely to prove useful in designing training courses.

Given these observations, an attempt was made to determine whether a variation on the cost accounting procedures suggested by Cronshaw and Alexander (1985) and Boudreau (1983a, 1983b), might be used to appraise the utility of alternative training interventions. To accomplish this, a general performance prediction model developed by Mumford, Weeks, Harding, and Fleishman (1988) describing the determinants of performance in Air Force resident technical training was employed, and a set of equations were constructed for determining the training performance costs implied by changes in training outcome variables specified in this model. In the present study, simulated changes were made in the variables determining performance, and the cost consequences of these changes were assessed.

Methods

Outcomes: To identify the major outcomes of Air Force resident technical training, a series of interviews were conducted at four major training centers. In these interviews various representatives of the groups responsible for the design and implementation of training programs were asked to identify the major outcomes of the technical training process and potential measures of these variables. On the

basis of this information, seven major outcomes of the technical training process were identified, including a) the assessed quality of academic performance; b) special instructional assistance time (SIA); c) number of academic counseling sessions; d) number of nonacademic counseling sessions; e) number of retraining hours or washback time; f) academic attrition, and g) nonacademic attrition.

Outcome costing: Having specified the major performance outcomes observed in Air Force resident technical training, the next major step entailed in this effort required the costs of differential performance on each of these criteria to be assessed. Because academic performance was held to be of interest only as it acted to determine other performance, relevant outcomes cost equations were not developed for this variable. The costs associated with performance on the remaining outcome variables were derived from Air Force cost accounting data applicable during 1983. Within this system, costs are defined in terms of three basic categories: direct costs (including officer, enlisted, and civilian pay as well as nonpersonnel costs of materials and supplies), indirect costs (staff benefits and material maintenance costs), and student costs (training pay, allowances, and travel). All cost estimates are derived through historic procedures, include appropriate discounting adjustments, and are phrased in terms of an individual trainee per unit time.

Elimination or attrition costs were held to include direct and indirect personnel costs and all student costs, including travel to training. The costs of outcomes involving increased training time (e.g., SIA time, academic and nonacademic counseling sessions of one hour, and retraining or washback time) were held to include the same elements except travel to training. This difference was based on the observation that the Air Force's travel investment is not lost if the individual graduates from training.

Given these cost figures per hour training time, it became possible to formulate equations for linking changes in outcomes to their associated costs. In the case of elimination, this was accomplished by assuming that elimination occurred half way through the course. Subsequently, half the course length was multiplied by the number of attritions and elimination costs to arrive at performance costs associated with academic and nonacademic attrition. The SIA time, washback time, academic and nonacademic counseling costs were calculated by determining the number of hours of extended training time obtained on each variable across all students and multiplying this value by the hourly training cost.

Modeling: Having identified the major performance outcomes of Air Force resident technical training, and developed procedures for appraising the costs associated with negative performance outcomes, it was then necessary to specify the variables influencing performance across training programs. To accomplish this, a series of group interviews were conducted at four major technical training centers in which key training personnel were asked to specify the student characteristics and course-content variables most likely to condition performance, the expected relationships among these variables, and potential measures of each construct. In all, 6 student characteristics and 16 course content variables were identified which appeared to influence performance across training programs.

To establish the relationship between training outcomes and these student characteristic and course content variables, a multivariate modeling effort was conducted. Initially, operational measures of the student characteristic, course-content, and training outcome variables were constructed and their reliability established. Subsequently, measures of these variables were obtained in 39 1983 training courses containing 5078 students. After establishing the relationships among these variables, a multivariate model was constructed which was intended to summarize the observed relationships in a manner consistent with the interview data. The ability of this model to replicate the observed structural relationships was then tested in a LISREL V analysis. This analysis yielded a residual term of .19 and a goodness of fit index of .59 indicating adequate fit to

the observed data. The multiple R's between .35 and .75 generated for predicting the 7 training outcome variables confirmed this conclusion. Further, these relationships appeared to be relatively stable since upon cross-validation in a sample of 9 additional courses containing 890 students, shrinkage of 10 to 15 points was obtained in attempts to predict observed training outcomes. The model developed and cross-validated in this effort is presented in Figure 1, and has been described elsewhere in greater detail (Mumford, Weeks, Harding, & Fleishman, 1988).

Simulations: To examine how changes in aggregate student characteristics and course-content variables would influence training outcomes and the associated performance costs, a series of simulations were conducted. Here 10 training courses representing a diverse set of fields were targeted for study. Subsequently, the aptitude, reading level, and academic achievement motivation as well as the occupational difficulty and subject matter difficulty variables were increased and decreased by one and two standard deviations. These simulated changes were made for each variable and for various combinations of the student characteristics and course content variables under consideration. The projected outcomes obtained in each of these simulations were then used in conjunction with the costing equations to generate expected performance costs.

Results

Table 1 presents the results obtained in these simulation runs for 2 of the 10 occupational specialties using 1983 flow figures and cost accounting data. Here costs have been aggregated across the 6 performance outcomes flowing from learning as indexed by the assessed quality of performance variable. In all cases the expected performance cost is presented on a per student basis and across all students in the course. Additionally, the percent change from baseline is presented.

The data presented in Table 1 leads to four conclusions. First, the costs and benefits associated with changes in student characteristics or course content are clearly conditioned by the cost of training, the length of the training program, and the number of students to be trained. Second, changes in both student characteristics and course-content can induce comparable savings in performance cost when both sets of variables have a significant impact on training outcomes. Third, the costs induced by changes in any one variable may be offset by appropriate adjustments in other variables. Fourth, a relatively small change in performance or individual performance costs can result in substantial gains and losses in lengthy, more expensive training courses involving a number of individuals.

Aside from these general conclusions, the results obtained in this simulation point to two other noteworthy conclusions. One of these pertains to the costs incurred by the various outcome variables. Of these variables, it was found that changes in attrition or retraining time typically had the largest impact on performance costs. Additionally, it should be noted that the percent change in costs obtained using this cost accounting approach was comparable to that obtained by Schmidt, Mack, and Hunter (1984) using supervisory judgments.

Discussion

The cost estimates produced in this modeling effort should not be taken to provide an accurate reflection of real training costs and benefits. One reason for making this statement is that the present study focused on performance costs in training and so ignored the costs and benefits associated with differential job performance. The other reason is that the intervention costs associated with "real world" changes in student characteristics and course-content were not considered herein.

Even bearing these caveats in mind, however, we believe that the present study has served to illustrate the potential value of applying utility analysis procedures in program evaluation efforts. As would be expected under standard utility assumptions, the costs and benefits incurred by changes in student

characteristics and course content are not uniform but depend on the number of individuals to be trained and the cost of the training program. Further, this study has shown how utility data can be used to integrate multiple qualitatively different outcomes and how these outcomes vary in relative significance.

The fact that training performance costs derived from cost accounting data yielded results comparable to those derived from other utility analysis procedures points to one way this might be accomplished. It has often proven difficult to evaluate the utility of alternative training interventions due to the lack of a cost relevant judgment base and adequate data on actual job performance. However, the present study shows that routinely available cost accounting data might be used to bridge this gap, particularly when the relative effectiveness of alternative interventions is of concern rather than the absolute value of a given intervention. It is hoped that the present effort, by demonstrating the feasibility of this approach, will stimulate further efforts along these lines.

References

- Boudreau, J.W. (1983a). Economic considerations in estimating the utility of human resources productivity improvement programs. Journal of Applied Psychology, 68, 396-407.
- Boudreau, J.W. (1983b). Effects of employee flows on utility analysis of human resources productivity improvement programs. Personnel Psychology, 36, 551-576.
- Cronshaw, J.P., & Alexander, C.E. (1985). One answer to the demand for accountability: Selection utility as a distinction between capital budgeting and utility. Organizational Behavior and Human Decision Processes, 35, 102-115.
- Goldstein, I.L. (1986). Training in organizations: Needs assessment, development, and evaluation. Monterey, CA: Brooks/Cole.
- Guion, R.L. (1965). Personnel testing. New York: McGraw-Hill.
- Kirkpatrick, D.C. (1959). Techniques for evaluating training programs. Training and Development Journal, 13, 3-9.
- Mumford, M.D., Weeks, J.L., Harding, F.D., & Fleishman, E.A. (1988). Relations between student characteristics, course content, and training outcomes: An integrative modeling effort. Journal of Applied Psychology, 73, 443-456.
- Schmidt, F.L., Hunter, J.E., & Pearlman, K. (1982). Assessing the economic impact of personnel programs on work force productivity. Personnel Psychology, 35, 333-347.
- Schmidt, F.L., Mack, M.J., & Hunter, J.E. (1984). Selection utility in the occupation of U.S. park ranger for three modes of test use. Journal of Applied Psychology, 69, 490-497.

Table 1

Total and Per Subject Costs Resulting From Simulated Changes

Course - Aircraft Environmental Systems Mechanic (M - 512)

	Degree of Change														
	-2	-1	0	1	2	2	2	2	2	2	2	2	2	2	2
Total Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals
VARIABLES IN SIMULATION ^a															
1	17050	320	+14	16037	301	+8	148812	280	0	137287	258	0	125763	236	-15
2	17858	337	+20	16656	313	+12	148812	280	0	131256	247	-12	113447	214	-20
3	17904	337	+20	16536	311	+11	148812	280	0	132257	249	-11	115619	217	-22
1+2	18666	354	+26	171648	323	+15	148812	280	0	126516	238	-15	104730	167	-26
1+2+3	20587	361	+35	179041	337	+20	148812	280	0	118649	223	-20	8966	169	-35
4	123195	221	18	135459	256	-9	148812	280	0	161666	375	+16	192765	362	+14
5	123845	227	14	127229	258	-8	148812	280	0	160269	375	+16	192765	362	+12
4+5	10017	188	22	124475	234	16	148812	280	0	173150	427	+52	246840	464	+32
4+5+2	154213	290	14	178550	376	+20	148812	280	0	227728	427	+52	246840	464	+66
4+5+3	13056	245	13	154709	240	-4	148812	280	0	203139	382	+36	227994	419	+49
4+5+4	70573	323	-52	94611	178	35	148812	280	0	142921	259	-4	162536	306	+9
4+5+5	63676	32	65	66444	126	55	148812	280	0	117746	213	-23	132750	250	-10

Notes

- 1 Attitude simulated changes
- 2 Reading level simulated changes
- 3 Achievement motivation simulated changes
- 4 Subject matter difficulty simulated changes
- 5 Occupational difficulty simulated changes
- 1+2 Attitude and reading level simulated changes
- 1+2+3 Attitude and reading level and achievement motivation simulated changes
- 4+5 Subject matter difficulty and occupational difficulty simulated changes
- 4+5+2 Subject matter difficulty and occupational difficulty simulated changes with two deviation decrease in attitude, reading level, and achievement motivation

Course - Jet Engine Mechanic (M - 306)

	Degree of Change														
	-2	-1	0	1	2	2	2	2	2	2	2	2	2	2	
Total Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	Totals	
VARIABLES IN SIMULATION ^a															
1	42745	270	+16	79323	259	+9	72594	237	0	67517	221	-7	67460	204	-14
2	87015	284	+19	80183	262	+10	72543	237	0	65002	212	-10	57412	188	-20
3	85038	278	+17	79750	261	+10	72593	237	0	65435	214	-10	58278	190	-19
1+2	91421	299	+26	82303	269	+13	72593	237	0	62833	205	-14	5327	174	-26
1+2+3	97130	317	+35	85720	280	+18	72593	237	0	59466	194	-18	46625	152	-35
4	64286	210	-11	66778	218	-8	72593	237	0	78008	256	+8	84223	264	+11
5	62385	204	-14	67484	221	-6	72593	237	0	77812	254	+6	82754	270	+14
4+5	54078	177	25	61674	202	-14	72593	237	0	83627	273	+14	94364	308	+29
4+5+2	78570	257	-8	86116	281	-18	72593	237	0	100069	353	+49	146276	388	+63
4+5+3	67205	220	+7	74801	244	-3	72593	237	0	96754	316	-32	107511	351	-48
4+5+4	41037	134	+43	48547	159	+32	72593	237	0	70500	230	+5	81257	266	-12
4+5+5	28897	94	+60	35817	117	+51	70543	237	0	57505	188	+26	66730	223	-5

4+5+1 Subject matter difficulty and occupational difficulty

simulated changes with one deviation decrease in

aptitude, reading level, and achievement

motivation

4+5+1+3 Subject matter difficulty and occupational difficulty

simulated changes with one deviation increase in

aptitude, reading level, and achievement

motivation

4+5+2+3 Subject matter difficulty and occupational difficulty

simulated changes with two deviation increase in

aptitude, reading level, and achievement

motivation

Note^a Total costs given nearly flowNote^b Total cost per studentNote^c Percent change in cost from baseline where + indicates

increased costs and - indicates decreased costs

THE EFFECTS OF A MATHEMATICS REFRESHER COURSE ON SCHOOL ATTRITION

Gary R. Bunde

Training Development Unit
Naval Technical Training Center, Corry Station
Pensacola, Florida 32511

PURPOSE

The purpose of this study was to determine the effects of a one week mathematics refresher course on specific unit test scores, final course average scores, and the overall academic attrition level of the ELTECH course. The ELTECH course is a basic electronics technology course taught at the Naval Electronic Warfare School, Naval Technical Training Center, Corry, Pensacola, Florida.

BACKGROUND

During the development of the ELTECH course, it appeared as if some students were having difficulty with mathematics in some units of the course. To reinforce the students' knowledge and skill of this subject, the Electronic Warfare School developed a mathematics screening test and a one week mathematics refresher course.

The mathematics screening test consists of sixty questions which cover, in part, addition, subtraction, multiplication, division, scientific notation, algebra, and basic trigonometry. Raw scores are converted to scaled scores which have a range of 1 to 100. Both the mathematics screening test and the mathematics refresher course were based on the type of mathematics used in the ELTECH course.

PROCEDURES

The sample of students for this study was divided into two groups. These groups were known as the Non-Math group and the Math Group.

1. Non-Math Group.

The Electronic Warfare School developed the mathematics screening test before the one week mathematics refresher course. The test was administered to a number of students from April 24, 1986 to 18 September, 1987, prior to their beginning the ELTECH course. These students were then immediately classed up into the ELTECH course. The school monitored the progress of these students, and variable data points were collected on them. For this study, these students (and subgroups of these students) were called the Non-Math Group.

2. Math Group

On 18 September, 1987, the school completed development of the mathematics refresher course, and continued to administer the mathematics screening test to all students before they started the ELTECH course. However, those students who scored 79 or less on this test were now sent to the one week mathematics refresher course. Those students who scored 80 or above were classed up immediately in the first available ELTECH class just as before. Those students who had to be sent through the mathematics refresher course were, upon its completion, retested with the same mathematics screening test as a posttest and then classed up in the ELTECH course.

This study collected variable data points on all students who entered the ELTECH course between 18 September, 1987 and 29 February, 1988. This group is called the Math Group. The same variable data points, plus posttest scores, were collected on the Math Group as were on the Non-Math Group.

3. Non-Math and Math Subgroups.

The Non-Math and Math Groups were subdivided as to whether they were electronic warfare personnel returning from the fleet (called EW personnel), or recent graduates from recruit training who would be receiving cryptologic technician maintenance instruction (called CT personnel). Students who scored under 80 on the mathematics screening pretest were called the "Under-80 (Non-Math or Math) Group". Students who graduated from the course were called "Graduates", and students who were dropped from the course for academic reasons were called "Attrites".

SAMPLE DESCRIPTION

	Non-Math Group		Math Group	
	Number	Percent	Number	Percent
Study Sample Size				
EW	36	24%	145	56%
CT	116	76%	113	44%
Total	152	100%	258	100%

Pretest scores under 80				
EW	28	78%	126	87%
CT	89	77%	92	81%
Total of Sample	117	77%	218	84%

Graduates				
EW	30	83%	136	94%
CT	34	29%	80	71%
Total of Sample	64	42%	216	85%

Attrites				
EW	6	17%	9	6%
CT	82	71%	33	29%
Total of Sample	88	58%	42	15%

VARIABLES

The variables that were used for this study are defined below.

1. AR Scores--Student Arithmetic Reasoning Test on the Armed Services Vocational Aptitude Battery (ASVAB) selection test. This test consists of solving arithmetic word problems.
2. MK Scores--Student Arithmetic Knowledge Test on the ASVAB selection test. This test consists of algebra, geometry, fractions, decimals, and exponents.
3. Pretest Scores---In this study pretest scores represent scores on the mathematics screening test administered to students for the first time prior to any type of mathematics training provided by the Naval Electronic Warfare School.
4. Posttest Scores--Student scores on the same mathematics screening test after they have completed the mathematics refresher course.
5. Unit 3 Scores--Scores on the Unit 3 (Basic Electricity and DC Circuits) final academic examination. This Unit and Unit 4 were considered by course developers as the best in-course measurement of mathematical skill.
6. Unit 4 Scores--Scores on the Unit 4 (Basic Electricity and AC Circuits.) final academic examination.
7. Final Scores---The student's final course average score on all units of the ELTECH course. This score is given as a final grade on the course, and identifies whether or not the student passed the course.

FINDINGS

1. Does the mathematics screening test measure mathematical abilities?

To answer this question, pretest and posttest scores of the sample were correlated with their AR and MK subtest scores. For all correlations, the Non-Math Group and the Math Group were first combined into one large group. This large group was then subdivided into the EW and CT groups with the same variables correlated.

All but two of the correlations among the pretest, posttest, AR, and MK tests were significantly positive. Of those that were significant, one was significant at or above .05, and the rest were significant at or above .01. The two correlations that were not significant were between the AR scores and the pretest scores

of the EW's, and the AR scores and the posttest scores of the CT's.

Primarily because of the significant correlations of the pretest and posttest scores with MK scores, it was concluded that the mathematics screening test was a good predictor of numerical mathematical problem solving, and is therefore a measure of mathematic abilities for the ELTECH students.

2. Given the fact that the mathematics screening test measures mathematical abilities, do students learn mathematics in the refresher course?

To measure learning, a T-test between means was run between the pretest scores and posttest scores of those students of the Math Group who were required to take the mathematics refresher course. (Because of the number of T-tests conducted for this study, a protection level was applied to all T-tests. In order for a calculated T value to be considered significant at or above the .05 level, it needed to be at or exceed the .002 level.) The mean score on the pretest was over 60 for the EW's and the CT's so that students did have some pre-knowledge of mathematics when they came into the school. However, after taking the one week mathematics refresher course, this mean test score increased by over 28 points for the EW students and over 24 points for CT students. This represents a mean pre to posttest gain of 47% for the EW students and 39% for CT students. Both increases were significant beyond the .05 level. Thus, even though students have some knowledge of mathematics coming in, the one week course does refresh this knowledge, and brings the students "back up to speed" on the subject.

3. Do students who take the mathematics refresher course do better on mathematics related test scores and final scores than comparable students who do not take the mathematics refresher course?

To answer this question, a comparison was made between the Unit 3, Unit 4, and Final scores of the Under-80 Non-Math Group and these same scores of the Under-80 Math Group.

Before this comparison could be made, however, the question had to be answered as to whether the Under-80 Non-Math Group and the Under-80 Math Group were comparable in the first place on variables on which the two groups should not differ significantly. This was done by comparing (by T-test) the EW's and CT's of the Under-80 Non-Math Group with the EW's and the CT's of the Under-80 Math Group on the variables of Pretest, AR, and MK scores.

There was no significant difference on Pretest scores, AR scores, and MK scores between EW's and CT's in the Under-80 Non-Math Group and the EW's and CT's in the Under-80 Math Group.

After determining that the Under-80 Non-Math Group and the Under-80 Math Group were equal on math related screening variables, comparisons were made between the two groups on their Unit 3 and Unit 4 academic test scores and their Final scores.

There was no significant difference between Unit 3 and Unit 4 scores of the EW Under-80 Non-Math Group and of the EW Under-80 Math Group. This observation must be tempered, however, by the fact that the number of EW students in the Under-80 Non-Math Group was small for Unit 3 (N=26) and for Unit 4 (N=22).

There was no significant difference between the Under-80 Non-Math Group and the Under-80 Math Group on CT Unit 3 scores. However, Unit 4 scores of the CT's on the Under-80 Math Group were significantly higher than these same scores of the Under-80 Non-Math Group.

With the possible exception of CT's on Unit 4, it does not appear that the mathematics refresher course improves the scores of the mathematics related tests of the ELTECH course.

There was no difference between the EW's or the CT's of the Under-80 Math Group and the EW's and CT's of the Under-80 Non-Math Group on their final score. Thus it would appear that the mathematics refresher course does not seem to improve Final scores.

4. Finally does the review of mathematics help those students complete the course, who would otherwise have dropped out of the course had they not had this review.

To answer this question utilizing the samples which made up the Math and Non-Math Groups of this study, a chi square test was run. This chi square test includes only students (EW and CT combined) who scored under 80 on the pretest. This test indicates that there was a significantly greater proportion of Graduates in the Math Group than there were in the Non-Math Group. This means that there was a significant drop in attrition between the two samples. The mathematics refresher course probably contributed to this drop in attrition, however, the cause of the drop can not be exclusively limited to the mathematics refresher course.

A measure of attrition that the Electronic Warfare School develops and maintains is an ELTECH Attrition Report. This is a weekly cumulative tally of the number of students who enter, graduate, and attrite from the course. According to this report, the academic drop rate was 2.1% for EW's and 27.4% for CT's at the end of FY 87.

The attrition rate of the Math Group in this study was 6% for the EW's and 29% for the CT's.

It therefore appears that, based on the two samples of students in this study, there was a significant reduction in attrition between the Non-Math Group and the Math Group. However, based on annual records maintained by the school, the attrition rate for the EW's has substantially increased from the previous fiscal year, while the CT's attrition rate has slightly decreased.

ADDITIONAL INVESTIGATION

As a further investigation, comparisons on the variables were made between students who academically attrited from the course and students who remained in the course and graduated. For all these comparisons (except posttest scores), the Non-Math Group and the Math Group were combined.

There was no significant difference between Graduates and Attrites on pretest scores, but Attrites scored significantly lower on posttest, Unit 3 and Unit 4 scores. There was no significant difference between Graduates and Attrites on MK scores, but students who academically dropped out of the course had significantly lower AR scores.

SUMMARY

Over 80% of the students who come into the ELTECH course score below 79 on the mathematics screening test, and therefore are deficient in the basic mathematics skills required to get through the ELTECH course. This review of mathematics does in fact refresh these skills and bring the students up to a higher and common level of skills and knowledge. However, with the possible exception of CT's on Unit 4, there does not appear to be any statistical evidence that the mathematics refresher course is helping students attain higher scores in the Unit 3 and Unit 4 academic tests, nor on the Final test.

Based on the sample used in this study, the mathematics course does significantly reduce attrition. However, based on records maintained by the school, the attrition rate for EW's is substantially higher, but the attrition rate for CT's is slightly lower. Students who attrite from the course have equivalent pretest scores as students who stay in the course, but then seem to immediately start to fail. The Attrites have a tendency for lower posttest scores than students who remain in the course, and continue to have lower mathematics related test scores in spite of the mathematics refresher training.

ACCURACY AND ADAPTABILITY: AN INVESTIGATION OF STANDARD DISTANCE ESTIMATION PROCEDURES

Mark A. Guadagnoli, Gene W. Fober, Pamela M. Terry and Willie R. Harden
U. S. Army Research Institute for the Behavioral and Social Sciences

A common technique used to move to desired locations in unfamiliar areas, often in the absence of a map, is dead reckoning. Dead reckoning is navigating by using a compass to maintain direction and distance estimation to maintain location along the desired line of travel. Research reported here investigated the effects of terrain and visibility conditions on the accuracy of distance estimation by pace counting.

Pace counting is the standard technique of distance estimation taught to Infantry soldiers. Soldiers are required to count each (or every other) step for given distances, usually in 100 m increments. Pace count, defined as number of steps per 100 m, is determined by dividing the number of steps by the number of increments paced. This count may then be used as a basis for future distance estimation. Because of limited training time, soldiers typically estimate their pace counts from only a 200-400 m course over a road or hard-packed trail during daylight. Therefore, there may not be an effective transfer from training to the "real world" because pace count may vary over changing terrain and visibility conditions.

U.S. Army Field Manual (FM) 21-26 indicates that soldiers should adjust their pace counts for changing conditions. However, early research in the area of distance estimation suggested the use of one pace count for all conditions (Powers, 1964). Thus, the impetus for the present investigation was to assess the accuracy of pace counting under varying conditions and to resolve the discrepancies between FM 21-26 and earlier research (Powers 1964). Informational results were also desired for possible use in developing pace count adjustment factors. Two experiments were designed to replicate conditions from Powers (1964). Pace counting in daylight and darkness was compared on-road in Experiment 1 and off-road in Experiment 2.

EXPERIMENT 1

Experiment 1 was designed to test changes in pace count over conditions of daylight and darkness along a road. Based on FM 21-26 it was predicted that pace count would increase during darkness because soldiers tend to take shorter steps in reduced visibility. Powers (1964) would predict that the variability between conditions should not be great enough to warrant an adjustment to soldiers' average pace counts.

Method

Subjects. Subjects were 20 male soldiers enrolled in One Station Unit training (OSUT) at Ft. Benning, Georgia. All soldiers had no previous Army experience with land navigation methods including pace counting procedures.

Test Site. The training and testing area was a 400 m section of dirt road at Ft. Benning, Georgia. Terrain was relatively flat with few elevation changes. It was chosen to replicate that of previous pace count research (Powers, 1964). Florescent markers placed along the road at varying distances guided soldiers along a predefined course. Intra-marker distances were varied (10 to 25 m) to prevent their use as distance cues.

Procedure. Soldiers were randomly assigned to one of two groups. Group 1 performed the experimental protocol during darkness on Day 1 and daylight on Day 2. Group 2 had the reciprocal protocol schedule. Prior to participation soldiers were assigned a subject number and given standard pace count instruction similar to that normally given in Basic Training. All soldiers determined their pace counts on a 200 m pace course during daylight. Immediately following this instruction, soldiers were given a 30-minute break. This time was necessary to await nightfall following daylight training for soldiers scheduled to pace during darkness. For consistency, this interval was maintained for all conditions.

Soldiers began the course individually at two-minute intervals upon the experimenter's instruction. They were required to pace a 730 m (365 m out and back) course along terrain like the practice terrain. The distance of 730 m was chosen because it was long enough to overcome early-course variability (Powers, 1964). (A rounded or commonly-used number of meters was not used for distance to discourage soldier guessing. Soldiers were informed of this.) Prior to starting the course, soldiers recorded their 100-meter pace counts. Upon course completion, they recorded the number of paces counted for the entire course.

Results and Discussion

For each soldier a ground distance estimate was computed by dividing his total reported paces by his reported 100-meter pace count. Mean errors in distance estimation were calculated as a function of group and visual conditions. A 2 X 2 (Group X Visual Condition) analysis of variance was conducted on the error data (difference of estimate from actual distance). The analysis was conducted for both directional error (DE) and absolute error (AE). Directional error scores signify both direction and magnitude of error. Absolute error scores signify magnitude of error only, disregarding the mathematical sign.

Directional Error (DE). The Group x Visual Condition interaction was not significant $F(1,36) < 1.0$, indicating a similar pattern of results for both groups unaffected by the order of performance. The group main effect also was not significant, $F(1,36) < 1.0$, indicating that groups did not differ in pace counting ability. Of prime importance to Experiment 1 was a significant main effect for visual condition, $F(1,36) = 8.60$, $p < .01$. Examination of the means (Table 1) revealed that soldiers tended to underestimate distance during daylight ($M = -16.5$ meters) and to overestimate distance during darkness ($M = 18.6$ meters).

Absolute Error (AE). AE means were calculated for group and visual condition and are presented in Table 1. Analysis of AE revealed no significant interaction $F(1,36) < 1.0$, and no main effect for group, $F(1,36) = 2.02$, $p > .05$. Unlike results for DE, there was no significant difference in estimation error for visual condition, $F(1,36) = 1.0$. The lack of a significant visual condition main effect for AE coupled with a significant main effect for DE indicates that the errors in distance estimation were directional. That is, soldiers consistently underestimated distance during daylight and consistently overestimated it during darkness.

Percent Error (PE). Powers (1964) had reported that a subject's pace count did not differ by more than 3 % for any of his terrain (on-/off-road) and visibility (day/night) conditions. Since design and procedures of the

current research differed from Powers', no directly comparable statistic was available. For conceptual comparison, error distance for each soldier was converted to percent of total course distance to assess differences among conditions in magnitude of estimation error. Means were calculated for group and visual conditions. Examination of the means (Table 1) revealed that the percentages of total distance in error were slightly larger than the 3% maximum variance in pace count reported by Powers (1964).

TABLE 1. Mean Distance Estimation Error by Group and Visual Condition, Experiment 1.

DAY			NIGHT	
GROUP 1	Directional Error	-18.7 m	Directional Error	13.2 m
	Absolute Error	22.7 m	Absolute Error	30.4 m
	Percent Error	3.1 %	Percent Error	5.0 %
GROUP 2	Directional Error	-14.3 m	Directional Error	23.9 m
	Absolute Error	36.7 m	Absolute Error	38.5 m
	Percent Error	4.2 %	Percent Error	5.3 %

Implications. From the data it appears that inaccuracies do occur when estimating distance by the present pace count procedure. When visibility was good (daylight), soldiers tended to slightly underestimate distance traveled. In poor visibility (darkness), soldiers tended to overestimate distance traveled. Increased uncertainty resulting from reduced visibility may make soldiers tend to move more tentatively, taking smaller steps (more paces per 100 m) than during daytime pace counting practice. Because more steps are taken to travel the same distance, the use of a single standard pace-to-distance formula results in an overestimation. This overestimation is more interesting when one considers that the tendency for error occurred under ideal conditions. That is, during darkness, soldiers followed alongside of florescent markers. In less ideal situations, they would set their own courses and pace unguided.

Importantly, these inaccuracies appear to be predictable in direction and therefore might be offset by adjustments in the pace count formula based upon information which could be accrued by individual soldiers through extensive practice or training time. However, research investigations such as this one may permit development of standard adjustment factors which may be used in lieu of extensive training.

Although there was a significant main effect of visual condition for DE, a 3- to 5-percent error in distance estimation hardly seems critical. Therefore, teaching only one pace count as Powers (1964) advocated seems to have some merit. However, the trend of errors in the present study may be important. Soldiers typically learn pace count on clear, flat terrain like that used in this experiment. However, in the field this skill must transfer to varied types of terrain. This added change in conditions could add to or compound errors. Thus, Experiment 2 was designed to replicate Experiment 1 using different terrain.

EXPERIMENT 2

Results of Experiment 1 revealed that soldiers tend to underestimate distance during daylight and overestimate distance during darkness. Although these results support the recommendation of FM 21-26, mean distance errors were only about 3- to 5 % of total distance. Experiment 2 was designed to assess the magnitude of errors when soldiers are required to pace count over more difficult terrain.

Method

Method and procedure for Experiment 2 was identical to that for Experiment 1 except that test terrain was wooded and hilly with varying elevation and surface conditions unlike the flat road. Eighteen OSUT soldiers learned to pace along a road during daylight and were tested under off-road conditions during daylight and darkness.

Results

Mean distance estimation errors were calculated as a function of group and visual condition. A 2 X 2 (Group x Visual Condition) analysis of variance was conducted on the error data for both DE and AE.

Directional Error (DE). The Group x Visual Condition interaction was not significant $F(1,32) < 1.0$ indicating no order effect. Also, there was no significant difference between groups, $F(1,32) = 1.74$, $p > .05$. As in Experiment 1, there was a significant DE main effect for visual condition, $F(1,32) = 12.49$, $p < .01$. Examination of the means (Table 2) revealed that soldiers tended to underestimate distance during daylight ($M = -8.3$ meters) and to overestimate distance during darkness ($M = 65.5$ meters). These results replicated Experiment 1 and visual inspection indicated a greater trend for overestimation during darkness.

TABLE 2. Mean Distance Estimation Error by Group and Visual Condition, Experiment 2.

DAY			NIGHT		
GROUP 1	Directional Error	12.7 m	Directional Error	72.0 m	
	Absolute Error	38.0 m	Absolute Error	100.2 m	
	Percent Error	5.2 %	Percent Error	13.7 %	
GROUP 2	Directional Error	-29.3 m	Directional Error	58.9 m	
	Absolute Error	44.2 m	Absolute Error	58.9 m	
	Percent Error	6.1 %	Percent Error	8.7 %	

Absolute Error (AE). Analysis of AE revealed no significant interaction effect $F(1,32) = 2.70$, $p > .05$. The main effect for group also was not significant, $F(1,32) = 1.47$, $p > .05$. The main effect for visual condition was significant, $F(1,32) = 7.05$, $p < .01$, indicating that the magnitude of the errors differed. Examination of the means (Table 2) revealed that the errors were much larger during darkness.

Percent Error (PE). Mean PE are presented in Table 2. Visual inspection of the darkness means indicated that magnitude of soldier errors ranged from 8.7- to 13.7 % of total distance. This exceeds the level that would be derived from the 3 % variance in pace count found by Powers (1964) and supports guidelines by FM 21-26 that pace-count adjustments are required.

Summary and Discussion.

The purpose of the present study was to investigate the existing discrepancy between the pace counting guidelines of Powers (1964) who advocated use of a single pace count for all conditions and FM 21-26 recommendation that soldiers adjust their pace count to changing conditions. Therefore, present research was designed to replicate some of Powers' (1964) conditions and examine guidance from FM 21-26.

In both experiments, soldiers determined their average 100 m pace counts on a 200 m out-and-back course along a stretch of dirt road. In both experiments soldiers used their counts to estimate 730 m courses during daylight and darkness. The course was on-road in Experiment 1 and off-road in Experiment 2. Results of both experiments revealed a significant difference in average directional error during daylight and darkness. Examination of the means revealed that soldiers tended to underestimate distance during daylight and overestimate distance during darkness. These results are consistent with FM 21-26 which states that soldiers paces will shorten when visibility is reduced. In addition, results of Experiment 2 showed a significant effect of visual condition for absolute error, which when coupled with the DE effect and PE larger than that from Experiment 1 indicate that errors may compound when soldiers are required to pace count in situations different from training. That is, when visibility only was considered, errors were smaller than when terrain conditions also differed.

One question that arises is whether these results are of practical significance. Powers (1964) advocated the use of only one pace count because he found that counts did not vary more than 3 % under differing conditions. Experiment 1 found similar percentages of error (3.1-5.3 %). In Experiment 2, however, mean percent error was 11.2 % for night pacing (both groups combined). This could be thought of as an error rate of 11 meters per hundred traveled, or 110 meters per kilometer traveled.

This level of error suggests that corrections are in order, either through training or during performance. It might be noted that a recent survey of Infantry platoon sergeants and leaders conducted by the Army's proponent agency for land navigation, the 29th Infantry Regiment, indicated that 42 % of soldiers coming from Initial Entry Training could not successfully perform the function of unit pace man. It is not known what level of accuracy was demanded, but present results suggest that one explanation for the failure of school-taught pace skills to translate to unit requirements may be that use of an unadjusted pace count in differing terrain and other conditions results in unacceptably large errors. If so, the solution is as mentioned above, development of individual adjustments from more training and practice or development of standard adjustment formulas should additional research continue to show error trends systematic enough to make this feasible.

Given that it might prove costly and difficult to reduce or contain the levels of distance estimation error through training, a second alternative would be to use pace count as a backup rather than a primary location

indicator whenever other cues are available. For example, when giving directions to the local supermarket, it is not most expedient to say "Turn right after traveling 2.3 miles." Instead, identifiable landmarks, e.g., "Turn right at the Gulf station." are provided. One might add that "The Gulf station is about 2 miles from here.", using distance estimation as a backup. Pace counting could be used in the same way. Soldiers could probably more accurately use landmarks from a map than estimated distances as primary location cues. That people tend to function more accurately using specific reference points rather than estimating distance is consistent with research using linear positioning tasks which shows subjects who use a counting (distance estimation) strategy to be less accurate than subjects who use reference points (e.g., location of body parts)(Reeve & Proctor, 1983). Since magnitude of estimation error is positively related to distance traveled, pace count would still remain the method of choice for distances too short to include identifiable landmarks.

Although not investigated here, pace counting is also cognitively inefficient because it requires constant attention. When an individual must do more than one task, performance suffers (Kahneman, 1973). Soldiers in field settings may have many tasks and, at minimum, must monitor their surroundings. Members of Army units also must avoid marking maps and thus must work from memorized routes and tactical plans. Rehearsal and execution from memory require much mental effort. Pace counting would surely be a performance detractor for soldiers navigating alone or where a unit pace man could not be spared.

Future research on distance estimation by pace counting should consider the cognitive demands, to include the mathematical functions of translation to ground distance and adjustment formulas. Future research on distance estimation error as a function of pacing conditions should include more conditions within the same repeated-measures design to allow for more comparisons within a single experiment and a better picture of interactive or independent operation of the various terrain and visibility condition factors. The range of individual pace-count variability within conditions might also be of interest. Meanwhile, further consideration should be given to navigational alternatives which minimize requirements for accuracy from methods and techniques which may be error-prone.

REFERENCES

- Kahneman, D. (1973). Attention and effort. Englewood Cliffs, NJ: Prentice-Hall.
- Powers, T. R. (1964). Advanced land navigation: Development and evaluation of a prototype program of instruction (Technical Report 89). Alexandria, VA: Human Resources Research Organization.
- Reeve, T. G., & Proctor, R. W. (1983). An empirical note on the role of verbal labels in motor short-term memory tasks. Journal of Motor Behavior, 15, 4, 386-393.
- U.S. Army. (1987). FM 21-26 Map Reading and Land Navigation.

**Intermediate Forward Test Equipment
Training Effectiveness Analysis
Kathy L. Nau and Gary G. Sarli
US Army TRADOC Analysis Command -
White Sands Missile Range
New Mexico**

In recent years, the Army has fielded new communication, aviation, missile, and other weapon systems that feature technologically advanced and sophisticated electronic components. To simplify fault diagnosis on these systems, special-purpose automatic test equipment (ATE) was developed for many new systems. Unfortunately, the resulting proliferation of ATE can increase manpower, training, and logistic support requirements.

One proposed solution to this problem is that the Army adopt a single, general-purpose ATE system. To be effective, this system must be highly mobile, modular, expandable, and adaptable to a large number of systems and electronics items. The Intermediate Forward Test Equipment (IFTE) system is being developed to meet these requirements.

A prototype IFTE system has been built by the Electronic Systems Division of Grumman Aerospace Corporation. All steps and instructions needed to measure performance of the electronic system being tested are part of a menu driven program that is displayed on a monitor during the fault isolation process. In addition, IFTE can detect and isolate its own faults and perform self-alignment and self-calibration procedures; this eliminates the need for ancillary test, measurement, and diagnostic equipment. The operator can configure IFTE to interface with and diagnose faults on a variety of electronic systems by inserting a cartridge containing appropriate system-specific software, the test program set.

IFTE consists of two subsystems: The Contact Test Set (CTS) and Base Station Test Facility (BSTF). The man-portable CTS, which interfaces with the system being supported, was designed for field troubleshooting and to augment system built-in test equipment. The CTS automatically tests, diagnoses, and verifies the operational status of the system and identifies failed line replaceable units (LRU) that will be removed and taken to a BSTF for further testing. The BSTF houses one or two base station test sets (BSTS) in a vehicle-mounted shelter. The operator uses a BSTS to diagnose and isolate faults within the LRU to the printed circuit board or module level (Green, 1986).

Disclaimer: The findings in this paper are not to be construed as an official Department of the Army position unless so designated by other official documentation.

The IFTE BSTF is to be staffed by personnel in the following Military Occupational Specialties (MOSs): Hawk Pulse Radar Repairer (24J), Communication - Electronic Radio Repairer (29E), Electronic Warfare/Intercept Tactical System Repairer (33T), Calibration Specialist (35H), Avionic System Repairer (35R), Automatic Test Equipment Operator/Maintainer (39B), and Fire Control System Repairer (45G).

Problem

Despite IFTE's advanced technology, its battlefield effectiveness depends upon the availability of qualified IFTE operators/maintainers. A Preliminary Training Effectiveness Analysis (Drackett, Southard, Risser, & Thorne, 1983) concluded that the major training issue associated with the system was the determination of training requirements for IFTE personnel.

Study objectives

The study had two objectives. The first was to compare the seven candidate MOSs for the BSTF operator/maintainer on aptitude, training, and job experience. The second was to identify the kind of additional training or training program modifications required for soldiers within each of the candidate MOSs to become IFTE BSTF operators/maintainers.

Limitations

CTS tasks were not examined in this study. TRAC-WSMR analysts had to develop the task list used in the study because the task list provided by Grumman (1986) lacked sufficient detail to support the planned analyses. Also, because IFTE fielding will not occur until the 1990s, TRAC-WSMR analysts did not have IFTE BSTF performance or training data to evaluate and could not conduct hands-on testing.

METHOD

Approach

For this study, a task is defined as a work activity that has a definitive purpose and that can be clearly defined as a segment of a job. Skill is defined as the ability, knowledge, or experience required to perform a task. The research strategy adopted for this study was based on the suppositions that tasks similar to those required for the operation and maintenance of the IFTE BSTF are currently performed by soldiers in the candidate repairer MOSs and that soldiers in the candidate MOSs already possess the skills required to perform many IFTE BSTF tasks. Therefore, data elicited from soldiers regarding MOS task and skill requirements may be used to make inferences about training requirements for parallel IFTE BSTF tasks. For further details, see Nau, Dowland, Sarli, Saia, and Martin (1988).

Sample

The candidate IFTE BSTF operator/maintainer study sample consisted of 357 Skill Level 1, 2, and 3 soldiers from candidate MOSs at 4 locations in the U.S. and 13 in Europe identified by the IFTE TRADOC Systems Manager, US Army Signal Center and Fort Gordon, Georgia. Additional information was obtained from 21 non-commissioned officers (NCOs) who were instructors and 34 E-6 first line supervisors for personnel in the candidate MOSs, and civilian IFTE subject matter experts (SMEs).

Data Collection Instruments

Instruments for Skill Level 1, 2, and 3 soldiers

Four data collection instruments were employed: A demographic questionnaire, a task survey, a written job test, and an electronics test.

The demographic questionnaire provided data on previous electronics training as well as demographic data such as age and time in the Army.

Seven MOS-specific task surveys determined how the soldiers had been trained to perform each task and how often they performed each task. The number of items on the surveys varied from 26 (MOS 33T) to 35 (MOS 35R), depending on the number of BSTF-related tasks their MOS performs.

Instructors from the MOS schools, with the assistance of TRAC-WSMR analysts, developed a written job test (WJT) for each MOS to evaluate soldier proficiency on IFTE BSTF tasks listed on the MOS task surveys. All WJT items were written in a multiple choice format with four alternatives per item.

TRAC-WSMR analysts, with the assistance of civilian SMEs, developed a 29-item, multiple choice electronics test to assess the level of electronics knowledge in the candidate MOSs (Leach, 1984; Miller, 1984). The items ranged from easy to very difficult, and were arranged in order of increasing difficulty in accordance with SME rankings.

Instruments for instructors and/or subject matter experts

The Electronic Principle Inventory (EPI) is an instrument developed by the United States Air Force for the specific purpose of determining principle skill and knowledge requirements for a job in electronics (AFPT 90-EPI-485). Sixty-three electronics subject areas are represented in this inventory. The EPI was used in the analysis to gain insight into what electronics skills soldiers in candidate MOSs use to perform their present MOS duties and what electronics skills IFTE BSTF operators/maintainers will need to perform their job. The response alternatives for all EPI items were simply yes and no.

The BSTF operator/maintainer task list contained 35 tasks, including tasks specific to the BSTF, and generic shelter, generator, and air conditioner tasks. This task list was used to determine which IFTE tasks are similar to tasks currently performed by soldiers in each MOS.

Instruments for supervisors

TRAC-WSMR analysts also developed a structured individual interview for first-level supervisors. The interview was designed to assess instructor familiarity with ATE and the impact of IFTE upon their units.

Additional data sources

The Enlisted Master File (EMF) contains information on soldiers in the candidate MOSs. These soldiers were considered to represent the study population. In order to develop population profiles for soldiers in each MOS, background information relating to variables such as age, rank, skill level, and time in the Army were obtained from the EMF. Armed Services Vocational Aptitude Battery (ASVAB) scores and Armed Forces Qualification Test (AFQT) scores obtained from the EMF were used to establish whether the MOS samples were representative of their respective populations. In addition, the Electronics (EL) score from the ASVAB was used to indicate basic electronics aptitude.

Procedure

Instructors from each MOS compared the IFTE BSTF task list to MOS task lists to determine whether MOS and IFTE task requirements were similar. They then used the EPI to indicate the skills and knowledge currently used by soldiers in the performance of their jobs. Next, the IFTE SMEs used the EPI and BSTF task list to determine skill and knowledge requirements of the IFTE BSTF operator/maintainer position.

The Skill Level 1, 2, and 3 soldiers were tested between February and August 1987. Groups of soldiers, ranging in size from 3 to 59, assembled in classroom-like environments where they completed the data collection instruments. TRAC-WSMR analysts conducted the individual interviews with repairer supervisors who were available during scheduled test sessions. Prior to beginning each interview, the supervisor was provided a written description of the IFTE system and the BSTF task list for reference.

RESULTS

Population and sample ASVAB and AFQT scores were compared using a multivariate analysis of variance (Harris, 1985). Because no significant between-group differences were found, MOS samples were considered to be representative of their respective populations.

Comparison of the IFTE and MOS task lists indicated that the majority of tasks needed to operate and maintain IFTE had counterparts in the candidate MOSs. Results from an analysis of the survey data reinforced this finding: Most of the MOS tasks identified by instructors as BSTF-related were performed by soldiers. Soldiers in the candidate MOSs possessed the majority of electronics skills required for the operation and maintenance of the IFTE BSTF. Additionally, soldiers in all MOSs except 29E were experienced with computer-assisted ATE.

WJT raw scores were converted to percent correct scores and Cronbach's Alpha reliability coefficients were computed for each test (Landy, 1985). Because only two of the reliability coefficients are within acceptable limits (0.70 or greater), WJT results should be interpreted with caution. Scores across MOSs ranged from 58 to 87%, suggesting that candidate MOS soldiers possess a basic level of proficiency on BSTF-related tasks.

Raw scores from the electronics test were converted to percent correct for analysis. The Cronbach's Alpha reliability coefficient was within acceptable limits. Although scores ranged from 35 to 54% correct, because IFTE SMEs stated that only a very basic knowledge of electronics is required to operate and maintain the IFTE BSTF, these scores do indicate that soldiers in all MOSs possess a basic level of electronics proficiency. This conclusion is supported by the soldiers' Electronic Composite scores (110 to 123) on the ASVAB.

Less than half of the NCOs interviewed supervised soldiers who operated and/or maintained ATE. Nevertheless, most supervisors anticipated some problems should the responsibilities of IFTE operation and maintenance be delegated to their units. Among the anticipated problems were skill loss, software inadequacy, software unavailability, and the need for familiarization training. Despite these problems, NCOs thought that IFTE could ultimately reduce troubleshooting time and decrease turnaround time. Because of the close similarity between IFTE and MOS task requirements and duties, 76% of supervisors stated that soldiers would need only familiarization training or no additional training.

CONCLUSIONS

The majority of tasks required to operate and maintain IFTE are generic tasks familiar to soldiers in the candidate MOSs. The soldiers possess most of the skills required for the operation and maintenance of the IFTE BSTF and are performing tasks similar to IFTE BSTF tasks. MOS 29E training will need to include exposure to and experience with computer-assisted ATE.

MOS 24J, 29E, 33T, 35H, 35R, 39B, and 45G soldiers could be trained as IFTE operators/maintainers without major training

program modifications, and IFTE BSTF operator/ maintainer training could be easily integrated into individual AIT courses.

REFERENCES

- Drackett, Maj. G. F., Southard, L. D., Risser, R. C., & Thorne, H. W. (March 1983). Direct Support Automatic Test Support System (DS ATSS) preliminary training effectiveness analysis (PTEA) (TRASANA TEA-11-83). White Sands Missile Range, New Mexico: U.S. Army TRADOC Systems Analysis Activity.
- Green, D. (Dec 1986). Minutes for Army Intermediate Forward Test Equipment (IFTE) ILSMT meeting number 4 and program review number 4 (IFTE 85-A0105-00). Bethpage, New York: Grumman Aerospace Corporation.
- Grumman Aerospace Corporation (1986). IFTE BSTF operator/ maintainer task list. Unpublished.
- Richard J. Harris, R. J. (1985). A Primer of Multivariate Statistics (2nd ed.). Orlando, Florida: Academic Press.
- Landy, F.J. (1985). Psychology of work behavior. Homewood, Illinois: Dorsey Press.
- Leach, D. P. (1984). Basic electric circuits (3rd ed.). New York: John Wiley & Sons.
- Miller, R. (1984). Electronics the easy way. New York: Barron's Educational Series.
- Nau, K., Dowland, Cpt R., Sarli, G. G., Saia, F. E., & Martin, M. (March 1988). Intermediate Forward Test Equipment Training Effectiveness Analysis (TRAC-WSMR-TEA-6-88). White Sands Missile Range, New Mexico: U.S. Army TRADOC Analysis Command.
- USAF Occupational Measurement Center, Occupational Survey Branch (1982). Electronic Principles Inventory (AFPT 90-EPI-485). Randolph AFB, Texas: U.S. Air Force.

Advanced On-the-Job Training: A Program Evaluation Model

Presenter:

Bernie Marrero, Ph.D.
Technical Advisor, AFHRL/OL-AK
Bergstrom AFB, Texas 78743-5000

Introduction

The Advanced On-the-Job Training System (AOTS) is a 'proof of concept' prototype. As a prototype, it is being field-tested in the operational setting to determine if automation can facilitate Air Force on-the-job training (AF OJT) management, delivery, and evaluation. The original mandate to design and develop AOTS is based on technical and inspection reports, noting deficiencies and needs in the labor and paper intensive AF OJT. These reports, in turn, led to the major directives that spawned AOTS.

The one year long system level test began on 1 August 1988. Given the decision-making emphasis on determining the merit and effectiveness of the AOTS, the use of program evaluation is a viable approach. Program evaluation has been viewed as a process in which data are obtained, analyzed, and synthesized into relevant information for decision-making (Borich, G. and Jemelka, R.P., 1982).

From the beginning, decisions have been an integral part of the AOTS. For example, decisions have been made on AF OJT needs, implementation strategies, feasibility, economy, and material resources. Similar to the design and development phases of the project, information will be needed for decision-making during the deployment phase. Ultimately, information collected during the entire project will contribute to the final decision on whether to implement the prototype AOTS.

Several authors in the field of program evaluation have taken the position that the role of evaluation is to provide decision makers with information (Alkin, 1969; Stufflebeam et. al., 1971, Borich, G. & Jemelka, R.P., 1982). Alkin (1969) views evaluation as the process of deciding on objectives, obtaining appropriate information to evaluate these objectives, and collecting and analyzing summative data useful to decision makers. The **Context Input Process Product (CIPP)** evaluation model is consistent with this decision making emphasis.

In the subsequent sections, the CIPP model will be discussed as a conceptual framework for evaluating the AOTS. This evaluation model will address the various phases of the AOTS as well as offer an integrated perspective for evaluating the entire system. The final section will include both advantages as well as disadvantages of this program evaluation model.

CIPP Program Evaluation Model

The Phi Delta Kappa Commission on Evaluation (Stufflebeam et. al., 1971) developed the CIPP model. This model divides evaluation into 4 distinct strategies- Context evaluation, Input evaluation, Process evaluation, and Product evaluation, thus the acronym CIPP

Each of these evaluation strategies address varying decision-making concerns. Context evaluation focuses on planning decisions to determine objectives. Input evaluation, on the other hand, serves structuring decisions by identifying and assessing design strategies that could be implemented to meet identified needs. Process evaluation involves implementation decisions to control the project intervention strategies. The last strategy, product evaluation, serves recycling decisions by relating outcome measures to context objectives, input, and process information.

The comprehensive scope and emphasis on decision making of the CIPP model is appropriate for the evaluation of the AOTS prototype. Each of the four strategies of this model involves an ongoing evaluation at different stages of the project that will culminate in the final phase of the deployment period. The objectives, method of obtaining information, and relationship of the information to decision making is addressed at each of the four levels of the evaluation model.

	CONTEXT EVALUATION	INPUT EVALUATION	PROCESS EVALUATION	PRODUCT EVALUATION
OBJECTIVE	DEFINE OPERATIONAL CONTEXT, NEEDS & PROBLEMS	IDENTIFY & ASSESS SYSTEM CAPABILITIES & INPUT STRATEGIES	PROCEDURAL EVENTS & ACTIVITIES	OUTCOME INFORMATION
METHOD	SUBSYSTEMS ACTUAL & INTENDED INPUTS OUTPUTS	DESCRIBE & ANALYZE RESOURCES, STRATEGIES & PROCEDURAL DESIGNS	MONITORING ACTIVITIES	COMPARING MEASUREMENTS & INTERPRETING OUTCOMES
RELATION TO DECISION MAKING	DECIDING SETTING NEEDS & OBJECTIVES	SELECTING SOURCES OF SUPPORT STRATEGIES DESIGN	IMPLEMENTING & REFINING PROGRAM	DECIDING TO CONTINUE TERMINATE MODIFY OR REFOCUS

CIPP EVALUATION MODEL

The figure above depicts the CIPP program evaluation model. In terms of the AOTS, the objective at the context level involves the evaluation of conventional OJT to identify needs and delineate problems underlying these needs. The method involves

collecting information on discrepancies between the intended outcome and the actual outcome of managing and delivering AF OJT. The final stage requires the formulation of decisions regarding goals associated with meeting the needs and objectives to solve the identified problems in OJT. In other words, making decisions for planning needed changes in the AF OJT.

Input evaluation is influenced by the decisions made during the Context evaluation. In order to address the objectives resulting from the context evaluation, decisions regarding input strategies and designs for implementing these strategies are made. The methodology utilized to make these decisions involve analyzing available human and material resources as well as considering the feasibility and economy of implementing the AOTS prototype. After formulating an implementation strategy, decisions are made to select sources for support, choose implementation strategies, and procedural designs. At this level, the AOTS was conceptualized, and preliminary designs were set into motion, leading to the development of the prototype. After the completion of the developmental phase, the AOTS would be deployed as a 'proof of concept' research study.

Once AOTS is deployed, the evaluation shifts to an overriding concern - monitoring the implementation of the prototype. Identifying procedural barriers and remaining alert to unanticipated ones are methods used during this phase of the evaluation. During the initial period of deployment, refinements in the program design and procedures can be made as a result of process evaluation decisions. After this initial period, however, process data is recorded and eventually used to enhance the interpretive value of the outcome measures.

Finally, the last phase of the model involves the evaluation of outcome measures. At the end of the deployment period, after utilizing the AOTS, users evaluate various facets of the system. Outcome information is interpreted in conjunction with information collected during the previous stages. That is, interpretation of outcome measures should be interdependent on the context, input, and process data. Decisions made during the previous phases of the evaluation should have a bearing on how the outcome data is interpreted.

At the end of the project cycle, the CIPP program evaluation should provide a dynamic baseline of information about the AOTS. In addition to the differential emphasis of the four strategies, information on the interrelationship of these strategies is obtained. Although the major emphasis of the evaluation will be on outcome information, the evaluation should provide additional information that may enhance our understanding of Air Force OJT, automated intervention strategies, and human factors related to changes in the operational setting.

Advantages and Disadvantages of the CIPP

The CIPP program evaluation model has advantages that must be weighed against disadvantages. Among the advantages are:

(1) The CIPP model provides the opportunity to review intermediate outcomes both during the evolution of the prototype and after the prototype has been deployed. For example, as AOTS evolves, information on the outcome of the design phase is available to make modifications. This advantage is particularly useful when a modification of the software in one of the AOTS functions is needed prior to deployment.

(2) The CIPP involves pre-established objectives to facilitate decision-making. These objectives provide implicit criteria for judging outcome measures.

While there are advantages to the model, some disadvantages are evident:

(1) The emphasis of the CIPP program evaluation model is on measurable short term objectives. This emphasis may have some inherent limitations. That is, short term objectives may not capture the benefits of the AOTS that may only be realized with measurable long term objectives.

(2) This model emphasizes the importance of determining measurable objectives and the criteria by which they are evaluated early in the project. However, there is minimal opportunity for the evaluator to determine the legitimacy and appropriateness of the objectives themselves. Once these objectives are established, they remain the same for the duration of the project. Given the 3 years since the formulation of the AOTS objectives, it is likely that situational events or developments may have required revisions of these initial objectives.

Summary

Despite the noted disadvantages of the CIPP program evaluation model, it provides a conceptual framework for collecting, organizing and interpreting field-based research. This model offers the opportunity to collect data during the evolution of the AOTS. It also allows the evaluator to focus on different types of information (e.g., context, process) to assess the system. However, its main emphasis is on the Gestalt of the system: the whole instead of its parts. Given these features, this model provides a comprehensive, integrated, and 'real world' approach to evaluate the AOTS prototype in the operational setting.

- Alkin, M.C., Evaluation Theory Development. In C.H. Weiss (Ed): *Evaluating Action Programs: Readings in Social Action and Education*. Boston: Allyn and Bacon Inc., 1972.
- Porich, G. and Jemelka, R. History and Growth of Program Evaluation. In *Programs and Systems: An Evaluation Perspective*. New York: Academic Press, 1982. (Chapter 1).
- Stufflebema, D.L., Foley, W.J., Guba, E.G., Hammond, H.D., Merryman, J., and Provus, M.M. *Educational Evaluation and Decision Making*. Itasca, Ill.: Peacock, 1971.

THE AIR FORCE TRAINING DECISION SYSTEM: R&D RESULTS

Winston R. Bennett, Chair
Air Force Human Resources Laboratory
Brooks Air Force Base, Texas

In September 1988, the Air Force Human Resources Laboratory (AFHRL) concluded a five-year exploratory Research and Development contract for a proof-of-concept Training Decisions System (TDS). TDS is a computer-based decision support system for use by Air Force Manpower, Personnel, and Training (MPT) decision makers in identifying the significant jobs of an Air Force Specialty (AFS), clarifying the training requirements of the major jobs in the AFS, and estimating the costs of specialty training (including typical classroom, field training detachment, correspondence, and on-the-job training programs). Initial development of the TDS has been completed and the system has undergone extensive testing and evaluation including sensitivity analysis and some data validation. Training has been provided to AFHRL and Air Training Command analysts and programmers. This panel provides an overview of the TDS, displays typical inputs to the system, data collection techniques and resulting data bases, job and training pattern simulations, cost and capacity calculations, and integrated specialty models. The potential uses of the TDS are discussed, as well as further development needed to exploit fully the possible utility of such a system.

OVERVIEW OF THE TRAINING DECISIONS SYSTEM: RESULTS AND PRODUCTS

David S. Vaughan
McDonnell Douglas Astronautics

For more than a decade, the Air Force has been using the Instructional System Development (ISD) model to guide the design of technical training for enlisted occupations and for support of new weapon system acquisition (see AFR 50-8). The ISD model requires a systems approach to training development, an approach aimed at providing "optimal training" for each specialty or weapon system. Many of the data elements required for the application of ISD, however, are not readily available in existing Air Force data bases. Further, not all decision algorithms suggested in the ISD model have been empirically validated nor are such algorithms equally applicable to all types of specialties. Decisions involving aircraft maintenance specialties may require an approach to training decision making that addresses issues which have a direct and immediate impact on combat sortie generation, whereas for a support specialty, the issues may involve primarily the cost per student or student flow limitations.

Recent developments in occupational analysis and training research, as well as in Air Force decision making processes (see AFR 50-8), have created new opportunities for optimizing training. Such developments include the recent emergence of Utilization and Training Workshops (U&TWs) and Training Planning Teams (TPTs) as the primary vehicles for making major training decisions. Such innovative procedural changes also make obvious a need for a technologically advanced data generation, analysis, and evaluation capability. To make good decisions about the training needed for an Air Force specialty or system, decision makers must be able to visualize and understand the jobs and training programs of the Air Force specialty (AFS) or weapon system under consideration and its technical training and Professional Military Education (PME) requirements, as well as the relative costs and payoffs of various training options. Such a "model" of a specialty provides a concise summary of the current status of the AFS, creates a common "language" for discussion or negotiation, and forms the baseline against which various alternative proposals can be evaluated.

To provide adequate support for such advanced training decision making, the Air Force Deputy Chief of Staff for Personnel, Education and Training (HQ USAF/DPPE) requested that the Air Force Human Resources Laboratory develop a computer-based Training Decisions System (TDS) to augment the Air Force ISD model. Such a system would generate necessary front-end training requirements data, validated decision algorithms, and procedures for improved interaction among training, personnel, and functional managers. The TDS would focus on supporting Air Force managers in making decisions as to the what, where, and when of the technical training (including the on-the-job training) required for a specialty (Ruck, 1982; Vaughan, Yadrick, Perrin, Cooley, Duntzman, Clark, & Rueter, 1984).

Background

Over several decades, the Air Force has evolved a task-based approach to determining technical training content and reviewing personnel

classification and utilization policies (Christal, 1974; Mitchell, 1988). As part of the occupational analysis (OA) process, tasks are defined by subject-matter experts (SMEs) of a specialty in their own technical terminology, working with analysts of the USAF Occupational Measurement Center, Randolph Air Force Base, Texas (see AFR 35-2). Several kinds of data on these tasks are collected from job incumbents and supervisors for use in reviewing training programs (see ATCR 52-22). Large samples of incumbents are asked to provide information about which tasks they perform in their present jobs and the relative amount of their job time spent performing such tasks. These data are used to examine the variety of specialized jobs within a specialty (occupation), to assess how jobs change at advanced skill levels, and to review official specialty descriptions and initial training programs (Christal & Weissmuller, 1988; Mitchell, Ruck, & Driskill, 1988).

By 1980, the determination of training setting was made at U&TWs, where trainers and training managers met with representatives from operational commands to negotiate training content and training setting (Mitchell, Sturdevant, Vaughan, & Rueter, 1987; see also ATCR 52-15). These conferences grew out of earlier procedures developed to bring initial skills technical training in line with initial job requirements ("HASTY GRAD" projects), while at the same time planning for those training requirements deferred to field training detachments (FTDs), mobile training teams (MTTs), or on-the-job training (OJT) programs (Ruck & Birdleough, 1977). Only minimal data were available for determining appropriate training settings for specialty tasks; thus, these decisions were, of necessity, based almost entirely upon the conferees' personal experience, or on known constraints at the resident training school. For these reasons, many of the decisions made in U&TWs cannot be consistently replicated. In addition, no formal evaluations or estimates were made of the impact of such decisions on personnel utilization, OJT costs, or mission performance (Ruck, 1982).

Technical Approach

In order to optimize training outcomes through providing the necessary instruction at an appropriate time in an airman's career and minimizing cost to the Air Force, a computer-assisted, data-based, decision support system was created in the TDS proof-of-concept project. The TDS, as developed, has three basic subsystems and a fourth integrating subsystem (see Figure 1; also see Vaughan, Yadrick, Perrin, Cooley, Duntelman, Clark, & Rueter, 1984).

The Task Characteristics Subsystem (TCS) provides methodologies and support for the creation of task training modules (TTMs) and procedures for collecting and analyzing training setting allocation data. TTMs are the prime building blocks for the TDS and are input to the other subsystems. The second subsystem, called the Field Utilization Subsystem (FUS), uses TTMs to describe present jobs and assignment patterns, to formulate alternative training and personnel assignment patterns, and to facilitate collection of managers' preferences among the alternatives patterns. These training and personnel assignment flow patterns become the prime focus of further data gathering and analysis in the TDS. The third subsystem, the Resource/Cost Subsystem (RCS), provides cost and capacity indicators for each task module for each training site. Such

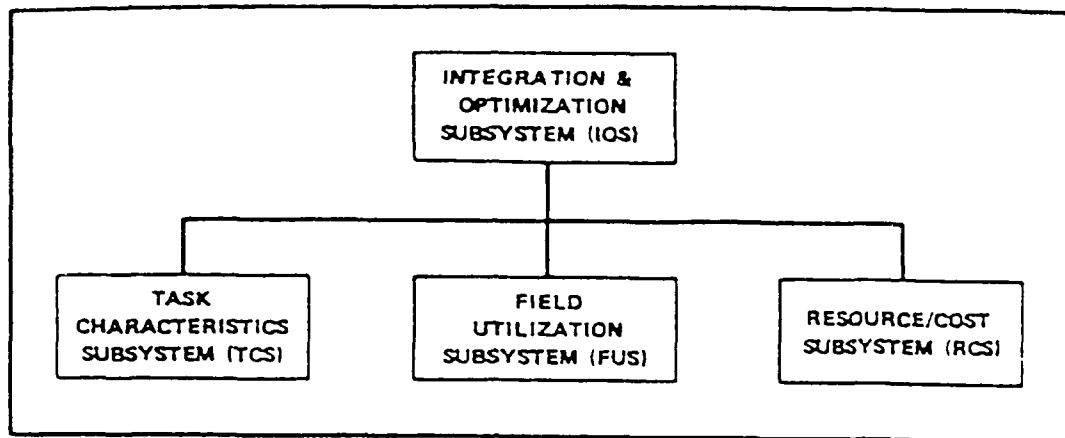


Figure 1. Major Subsystems of the Training Decisions System

cost and capacity functions are an important input to the training optimization routines performed in the Integration/Optimization Subsystem (IOS). The IOS is designed to help Air Force managers optimize training decisions using a number of objective functions. It allows changes in functions and constraints and generates management-oriented decision aids as well as more detailed reports to be staffed throughout the Air Force to facilitate more coordinated, informed training decisions.

The initial development of the TDS has been completed. We will provide an overview of each of the TDS subsystems in the papers which follow. Before we provide such details, I would like to provide you with examples of TDS products and outcomes, to better communicate the end objectives and capabilities of the system.

The approach used in TDS is to model each problem or potential solution as an independent Utilization and Training patterns. Once a problem or potential solution has been identified, it is expressed in terms or quantities of the TCS; i.e., task modules, training settings, and allocation curves. The FUS modeling capability can be used to translate the potential solution into modifications of the AFS data base; that is, as an alternative to the current U&T files (see Figure 2). FUS and RCS software are employed to generate products for each alternative (potential solution) and results, in the form of training costs and capacities, are compared to baseline data (current U&T costs and capacities) by Air Force decision makers.

This approach is a very flexible and powerful tool for Air Force managers and decision makers to use to examine the possible consequences of their decisions on total training costs for a specialty and on the constraints to the capacity of organizational units to conduct on-the-job training (OJT). This type of decision support has not been available before the development of the TDS, yet should prove to be highly valuable to various types of Air Force decision making bodies, such as Utilization & Training Workshops (U&TW), and Training Planning Teams (TPTs).

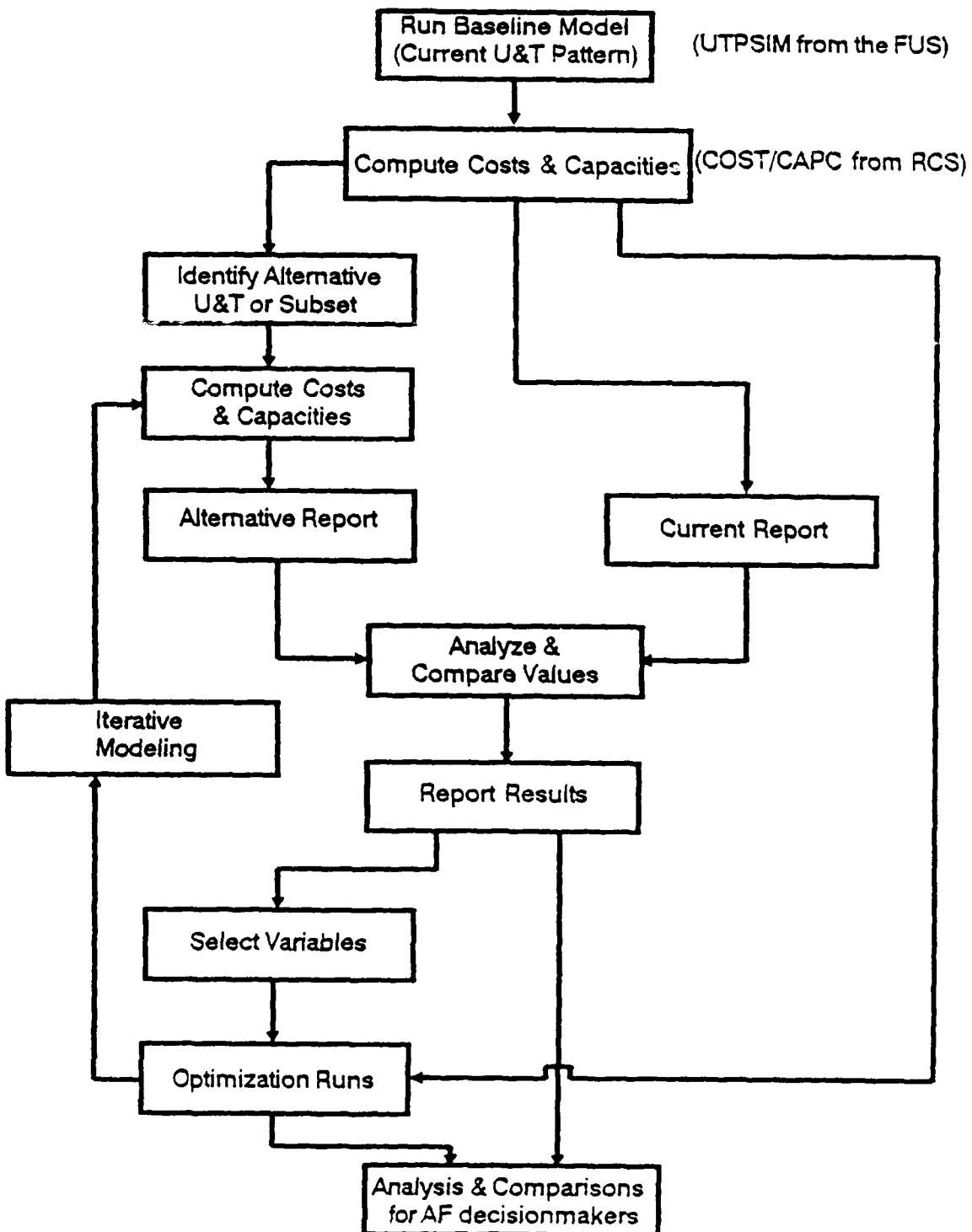


Figure 2. Relationship of Modeling and Optimization Functions of the TDS.

An example may serve to illustrate how potential changes in training can be evaluated. The problem might be a resource constraint in conducting OJT in some units, such as when not enough hours are available for training a specialized piece of test equipment--a Weapons Release Control System (WRCS) Analyzer in AFS 328X4, Radar and Inertial Navigation Systems Maintenance.

One approach to the problem would be to move training of that equipment from an OJT setting to some formal course, such as the basic resident course at the technical training center (TTC) or to a field training detachment (FTD). The first of these possibilities can be modeled in the TDS by adding enough hours to the resident course to achieve the required proficiency (as indicated by the Allocation Curve for the WRCS TIM). A second model would be to do the same for the FTD.

The results of these analyses are shown in Figure 3, along with data from the current U&T pattern as a baseline for comparison.

	<u>CURRENT</u>	<u>ABR COURSE</u>	<u>FTD COURSE</u>
ABR COURSE COSTS	\$1,676,352	\$1,696,406	\$1,676,352
FTD COURSE COSTS	\$ 45,647	\$ 45,647	\$ 53,495
TOTAL COURSES	\$2,724,296	\$2,744,350	\$2,732,145
OJT COSTS	\$5,096,500	\$5,076,792	\$5,095,981
OJT CAPACITY	EXCEEDED	NOT EXCEEDED	EXCEEDED

Figure 3. Comparison of AFS 328X4 U&T Patterns Involving Movement of WRCS Analyzer Training.

It should be noted that the "EXCEEDED" under the current U&T pattern indicates that some resource constraint exists (in this case, an equipment availability constraint) so that not all of the required training is being provided. Note also that the proposed solution of adding the required WRCS Analyzer training to the FTD does not solve the problem; OJT capacity is still exceeded.

Moving the training to the resident (ABR = Airman Basic Resident) course appears to solve the constraint problem, but at an additional ABR cost of approximately \$20,000. However, there is also a reduction in OJT costs of about the same amount (since this training is no longer provided in the field). Thus, this solution seems to be a viable option which resolves the problem at no major increase in total AFS training costs.

This particular example involved proposed changes to training programs but the same type of analysis can be done for changes which might involve redefining jobs, or changing the flow of personnel through particular training programs and jobs. In any of these situations, the TDS modeling and costing approach is able to project the relative consequences (constraints and costs of the new pattern) and provide Air Force decision makers with objective data on which to base critical training decisions.

References

- Air Force Regulation 35-2 (1982, 23 July). Occupational analysis. Washington, DC: Headquarters, United States Air Force.
- Air Force Regulation 50-8 (1984, 6 August). Policy & guidance for instructional systems development. Washington, DC: Headquarters, United States Air Force.
- Air Training Command Regulation 52-15 (1982, 24 September). Career field utilization and training workshops (U&TW). Randolph AFB, TX: Headquarters, Air Training Command.
- Air Training Command Regulation 52-22 (1981, 16 October). Occupational analysis program (corrected copy). Randolph AFB, TX: Headquarters, Air Training Command.
- Christal, R.E. (1974). The United States Air Force occupational research project (AFHRL-TR-73-75, AD-774 574). Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory.
- Christal, R.E., & Weissmuller, J.J. (1988). Job-task inventory analysis. In S. Gael (Ed), Job analysis handbook for business, industry, and government. New York: John Wiley and Sons, Inc. (Chapter 9.3).
- Mitchell, J.L. (1988). History of job analysis in military organizations. In S. Gael (Ed), Job analysis handbook for business, industry, and government. New York: John Wiley and Sons, Inc. (Chapter 1.3).
- Mitchell, J.L., Ruck, H.W., & Driskill, W.E. (1988). Task-based training program development. In S. Gael (Ed), Job analysis handbook for business, industry, and government. New York: John Wiley and Sons, Inc. (Chapter 3.2).
- Mitchell, J.L., Sturdevant, W.A., Vaughan, D.S., & Rueter, F.H. (1987). Training decisions system: Information gathering technical paper (draft Technical Report, CDRL 23). Brooks AFB, TX: Prepared for the Training Systems Division, Air Force Human Resources Laboratory.
- Ruck, H.W. (1982, February). Research and development of a training decisions system. Proceedings of the Society for Applied Learning Technology. Orlando, FL.
- Ruck, H.W., & Birdlebough, M.W. (1977). An innovation in identifying Air Force quantitative training requirements. Proceedings of the 19th Annual Conference of the Military Testing Association. San Antonio, TX: Air Force Human Resources Laboratory and the USAF Occupational Measurement Center.
- Vaughan, D. S., Yadrick, R. M., Perrin, B. M., Cooley, P. C., Duntelman, G. H., Clark, B. L., & Rueter, F. H. (1984, August). Training decisions system preliminary design (draft Technical Report, CDRL 21). Brooks AFB, TX: Prepared for the Manpower and Personnel Division, Air Force Human Resources Laboratory

THE TASK CHARACTERISTICS SUBSYSTEM:
ALLOCATING TASK MODULES TO TRAINING SETTINGS

Bruce M. Perrin & J. R. Knight
McDonnell Douglas Astronautics

The Training Decisions System (TDS) is a computer-assisted, data-based decision support system intended to help optimize Air Force training and facilitate coordinated, informed training decision making. The TDS consists of three basic subsystems and a fourth integrating subsystem. The Task Characteristics Subsystem (TCS) provides data on training setting allocations and on task training modules (TTMs). TTMs are the basic building blocks of the TDS, in that all training courses and jobs are described in terms of TTMs, and training costs are determined using TTM-based procedures. The Field Utilization Subsystem (FUS) provides information for defining training and job assignment patterns, as well as information on management preferences for the current and plausible alternative approaches to training, assigning, and using airmen in a particular Air Force specialty (AFS) over the span of their Air Force careers. The Resource/Cost Subsystem (RCS) provides estimates of training resource requirements, costs, and capacities at various training sites in an AFS. Finally, the Integration/Optimization Subsystem (IOS) integrates the other three subsystems into a single model for evaluating training cost and capacity impacts of training and job structure changes in an AFS, in order to generate management-oriented decision aids and detailed reports. This paper provides a detailed description of the TCS.

TTM Construction Component

The Task Characteristics Subsystem (TCS) provides methodologies and support for the creation of task training modules (TTMs) and procedures for collecting and analyzing training setting allocation data. TTMs are the prime building blocks for the TDS and are used as a basis for collecting data in the other TDS subsystems. The use of TTMs solves a number of problems associated with task-level data (see Perrin, Knight, Mitchell, Vaughan, & Yadrick, 1987).

The Air Force presently makes substantial use of task data for training decisions and other purposes; one problem is that many tasks tend to share common skills and knowledges, and thus, normal task analysis procedures typically produce very repetitive data. TTM-level data, however, reflect shared skills and knowledges, thus reducing the possibility of overestimating training requirements.

A second problem in using task data is the fact that each specialty involves 300 to 2,000+ tasks, far too many for managers to process in a typical Utilization and Training Workshop (U & TW). Indeed, U & TW participants generally focus on review of the Specialty Training Standard (STS), leaving detailed review of tasks to occupational analysts and training developers. Task analysis of all the tasks of a specialty is a very time-consuming, labor-intensive, expensive process. In addition, various types of tasks may require different types of analysis (DeVries, Eschenbrenner, & Ruck, 1980; Eschenbrenner, DeVries Miller, & Ruck, 1980).

The fact is that the Air Force cannot afford the manpower and expense of a detailed task analysis for every task of every specialty. Therefore, a procedure is needed to group or cluster tasks which share common skills and knowledges; i.e., those tasks which could be trained together most efficiently. In the TDS, individual tasks are grouped into clusters of related tasks in the TTM construction component.

Two approaches to the problem of constructing TTMs were evaluated empirically in the TDS R & D effort (Perrin, Vaughan, Yadrick, Mitchell, & Knight, 1986). One was a judgmental approach, using the expertise of subject-matter experts (SMEs); a second approach used data from the most recent occupational survey to cluster tasks statistically.

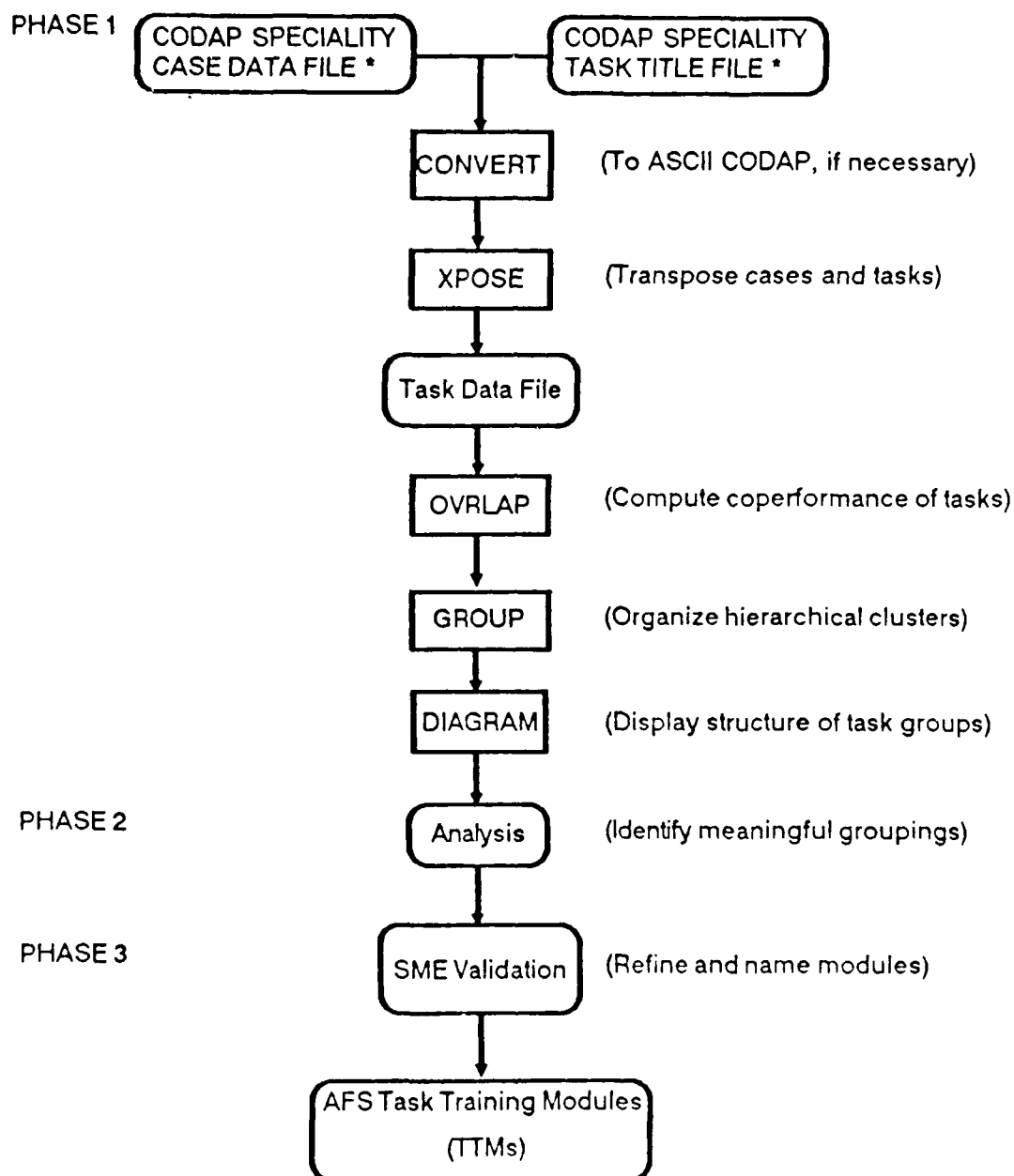
In the judgmental approach, SMEs were asked to sort the tasks of their specialty into subjective categories; i.e., to group those tasks which should be trained together. Presumably, the resulting task groupings would: (a) contain tasks which share similar underlying skills or knowledges, and (b) include tasks performed in the same job. It is a "given" that the same task might appear in more than one job. Results from the attempts to apply this judgmental approach in the initial two specialties (AFS 328X4 and 811XX) were not encouraging, since each group of SMEs tended to use unique conceptual approaches depending on their backgrounds and present assignments. Even though a fair consensus was achieved through extended negotiation, a complete consensus was not attained. Attempts to refine the clusters with new reviewers resulted in only minor modifications based on their unique perspectives of their AFS.

In the second approach, multivariate statistical techniques were employed in an iterative manner with a variety of input data, such as the probability of co-performance of tasks calculated from available occupational survey report (OSR) data, common equipment usage, and skill-level or grade information. Initial trials indicated that the latter variables added little or no refinement to statistical clusters formed using only a measure of task co-performance. Advanced work focused on use of Comprehensive Occupational Data Analysis Programs (CODAP) hierarchical clustering of tasks and interpretation of meaningful clusters of tasks as TTMs. It was found that the help of SMEs was critical to naming the modules, assessing their significance, and allocating isolated tasks which did not statistically cluster.

The final recommended procedure was a combined approach that utilized statistical co-performance clustering to form initial groupings of tasks, having an analyst identify task groupings which appeared meaningful, and then having SMEs title and refine the task groupings into TTMs (see Figure 1). This approach saves time for both analysts and SMEs and provides a structured focus for their efforts. Detailed task analysis of the tasks of a sample of TTMs for two AFSS indicated the tasks in TTMs formed through this process did indeed share common skill and knowledge requirements (Perrin et al., 1986).

Training Setting Allocation Component

Once TTMs are finalized, survey instruments are developed with which to gather information as to how TTMs are and should be allocated to



* CODAP AFS study files available in the AFHRL archives (historical) or the USAFOMC (current studies).

Figure 1. Development of Task Training Modules from Occupational Survey Data using ASCII CODAP co-performance clustering (see Perrin et al., 1986; see also Perrin et al., 1987, Appendix A)

various training settings. Groups of senior technicians in the specialty, who are thoroughly familiar with the work of the specialty, estimate how much training time is currently devoted to reach minimum required proficiency for the various groups of tasks for the following training settings: classroom, correspondence courses such as career development courses (CDCs), hands-on training (FTDs, MTTs, etc.), supervised hands-on experience on the job (OJT), and other programs. These raters are also asked to provide training time estimates for "ideal" training (i.e., the most effective mix of types of training). Finally, the raters are asked how long it would take in each setting to train the TTM, if the training was provided in only that setting (only classroom training, only correspondence course training, etc.). Of course, this "maximum effective training" may not always yield full proficiency; it may not be possible to fully train a TTM in one or more of the settings alone. In these cases, the respondents were to indicate the proficiency level reached in the setting as a percent of full proficiency. Thus, each SME provided six allocation judgments for each TTM: four that related the maximum training time in each of four settings to the proficiency level reached, one for the current allocation, and one for the most preferred allocation of training.

Difficult issues such as proficiency measurement and the description of partial allocations of training to different training settings were generally resolved to the satisfaction of the many SMEs who were involved in the development of these procedures. For TDS, proficiency was defined as a percentage of the training needed by an average individual to reach the minimum required standard for each TTM (the "go/no go" level of OJT = 100% proficiency). SMEs generally understood and were able to estimate degrees of proficiency expressed in this way. SMEs are also able to reliably and consistently describe the current training pattern in terms of how training in each type of setting contributes incrementally to the attainment of full proficiency (i.e., the partial proficiency achieved).

It was hypothesized that proficiency gain from training in a setting would be greatest initially and would decline as more training was provided in that setting. Eventually, there would be no more gain from providing training. Thus, the predicted relationship between proficiency and time in a training setting is that of initial gain followed by proficiency leveling-off, a negatively accelerated curve. This general set of relationships is depicted in Figure 2. These curves can be modeled by the following polynomial regression equation:

$$\begin{aligned} \text{Proficiency}' = & a * \text{class-hours} - b * \text{class-hours}^{**2} + \\ & c * \text{self-study-hours} - d * \text{self-study-hours}^{**2} + \\ & e * \text{field-training-hours} - f * \text{field-training-hours}^{**2} + \\ & g * \text{work-hours} - h * \text{work-hours}^{**2}, \end{aligned}$$

where "a" through "h" are coefficients to be estimated by multiple regression, **2 indicates squaring, and the regression equation is constrained to pass through the origin (there is no constant for the Y intercept).

This model involves specific hypotheses about the nature of the relationship between training hours in each setting and proficiency. Specifically, controlling for training in each of the other training settings, the first-order parameter is specified to be positive and the

second-order parameter is negative, yielding the predicted negatively accelerated curve.

Across the four AFSs studied during TDS development, this statistical model was strongly supported. Statistical estimates consistent with the polynomial regression equation were found in well over 90% of the TTM allocation curves in all four specialties. The overall fit of the polynomial regression model was found to be quite good, averaging over 65% (multiple R squared) in AFSs 423X1 and 305X4. The additional variance explained by second-order terms in the allocation equations was substantial (approximately 15% increase in R squared for the specialties), indicating that simple linear functions are not sufficient to describe proficiency gains from training in each setting (see Figure 3 for an example of the allocation curves for an AFS 328X4 TTM). A curvilinear model is much more descriptive of these relationships (Perrin, Knight, Mitchell, Vaughan, & Yadrick, 1988).

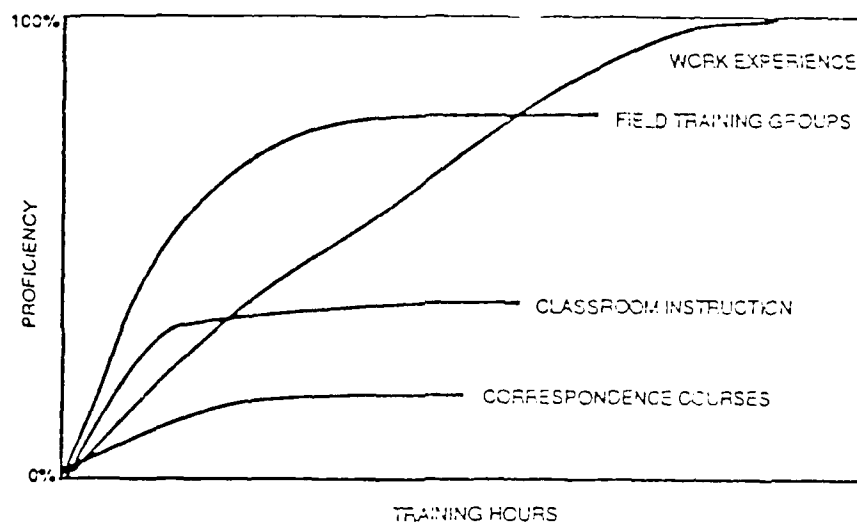


Figure 2. Hypothesized Relationship Between Hours of Training in a Setting and Proficiency Gain (from Perrin et al., 1988:29).

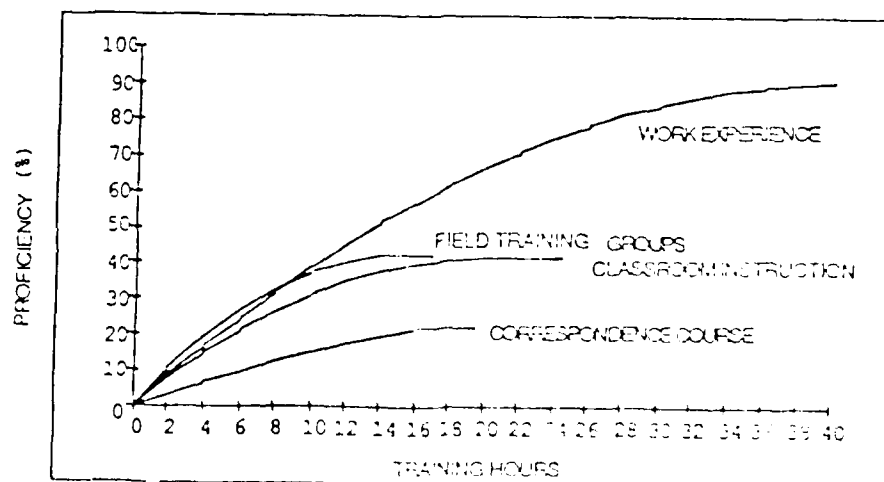


Figure 3. Example Allocation Curves for AFS 328X4 TTM 34, Doppler Sensor Control Boxes.

Allocation curves for all TTMs of a specialty derived through this survey approach give the TDS maximum flexibility in considering different ways of dividing training among training settings, as well as identifying the limits of each type of training. This capability serves as one basis for developing and evaluating alternative patterns of training. Thus, the allocation curves are a very significant part of the overall TDS design; their development and validation represent a substantial advance in training decisions technology.

Relationship of the TCS to Other Subsystems

The set of TTMs developed and validated for a specialty represent the major building blocks for the TDS. The TTMs are major inputs to the FUS and serve as descriptors for jobs and training states (courses, OJT programs, etc.). Thus, training and job content share a common set of terms. The TTMs also serve as a foundation for the RCS in that information about training resources required and resource availability are collected on a TTM-by-TTM basis.

References

- DeVries, P.B. Jr., Eschenbrenner, A.J., Jr., & Ruck, H.W. (1980, July). Task analysis handbook (AFHRL-TR-79-45[II], AD-A087 711). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Eschenbrenner, A.J., Jr., DeVries, P.B., Jr., Miller, J.T., & Ruck, H.W. (1980, July). Methods for collecting and analyzing task analysis data (AFHRL-TR-79-45 [I], AD-A087 710). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Perrin, B.M., Knight, J.R., Mitchell, J.L., Vaughan, D.S., & Yadrick, R.M. (1987, September; Revised January 1988). Task characteristics subsystem: administrative report (draft Technical Report, CDRL 10). Brooks AFB, TX: Prepared for the Training Systems Division, Air Force Human Resources Laboratory.
- Perrin, B.M., Knight, J.R., Mitchell, J.L., Vaughan, D.S., & Yadrick, R.M. (1988, September). Training decisions system: Development of the task characteristics subsystem (AFHRL-TR-88-15). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Perrin, B.M., Vaughan, D.S., Yadrick, R.M., Mitchell, J.L., & Knight, J.R. (1986, 7 February). Development of task clustering procedures (draft Technical Report, CDRL 7B). Brooks AFB, TX: Prepared for the Manpower and Personnel Division, Air Force Human Resources Laboratory.

THE FIELD UTILIZATION SUBSYSTEM:
JOB AND TRAINING PATTERN SIMULATIONS

J. L. Mitchell & R. M. Yadrick
McDonnell Douglas Astronautics

The Training Decisions System (TDS) is a computer-assisted, data-based decision support system intended to help optimize Air Force training and facilitate coordinated, informed training decision making. The TDS consists of three basic subsystems and a fourth integrating subsystem. The Task Characteristics Subsystem (TCS) provides data on training setting allocations and on task training modules (TTMs). TTMs are the basic building blocks of the TDS, in that all training courses and jobs are described in terms of TTMs, and training costs are determined using TIM-based procedures. The Field Utilization Subsystem (FUS) provides information for defining training and job assignment patterns, as well as information on management preferences for the current and plausible alternative approaches to training, assigning, and using airmen in a particular Air Force specialty (AFS) over the span of their Air Force careers. The Resource/Cost Subsystem (RCS) provides estimates of training resource requirements, costs, and capacities at various training sites in an AFS. Finally, the Integration/Optimization Subsystem (IOS) integrates the other three subsystems into a single model for evaluating training cost and capacity impacts of training and job structure changes in an AFS, in order to generate management-oriented decision aids and detailed reports. This paper provides a detailed description of the FUS.

Current Utilization and Training (U&T) Pattern Component

A description of the current U&T pattern for a given specialty is a necessary starting point both for understanding the current situation and for developing possible management choices. Appropriate data must be synthesized from a variety of sources--most notably from occupational survey (OS) data, the Uniform Airman Record (UAR), the Pipeline Management System (PMS), AFP 50-5, TDS surveys, and informal interaction with functional managers, training managers, and field representatives--to gain a complete picture of the present AFS training and personnel assignment flows. Such information must be summarized and displayed to Air Force managers in a simple, effective format that provides a brief yet comprehensive picture of the specialty.

A key element in defining a career pattern is the identification of the jobs within a specialty. In occupational analysis (OA) terminology "job" refers to a group of related positions in which many of the same tasks are performed; "position" refers to a unique set of tasks performed by one person (Shartle 1959). Each AFS includes a number of jobs that vary in content (tasks performed) according to the organizational level, unit mission, equipment operated or maintained, level of experience of personnel, and a number of other interrelated factors (Driskill & Mitchell, 1979).

For the TDS, the major job types were identified using standard OA methods for the analysis of data collected from job incumbents (Archer, 1966; Christal, 1974; Morsh, 1964). Using two test specialties, several attempts were made to create job clusters based on type of weapon system

or equipment maintained, on average grade of personnel performing tasks, and on other potentially relevant factors. Such statistical weighting procedures did not provide meaningful or clearly improved job types compared to those identified in the original OA job typing (Yadrick, Knight, Mitchell, Vaughan, & Perrin, 1988).

In addition, first enlistment jobs (1 to 48 months Total Active Federal Military Service or TAFMS) were contrasted with career (49+ TAFMS) jobs: however, no major differences, other than the expected amount of supervisory-type work, were identified. Thus, the OA job types appear to be a realistic foundation for describing the current U&T pattern for Air Force enlisted specialties, although issues remain regarding the level of analysis most appropriate for the TDS (Yadrick et al., 1988). Advanced job typing procedures are currently being developed and tested (Phalen, Staley, & Mitchell, 1987) in an effort to make the process more efficient.

For TDS, OSR data are reprocessed to create more concise job descriptions based on the TTMs of the specialty from the TCS. The CODAP set of MODULE programs (Module Title File, MODSET, PRIMOD) is used to create a new data base. Four indices are used to characterize the Job-TTM association:

1. the sum, across people and tasks, of the percent time spent performing the tasks of a TTM;
2. a running cumulation of this summed percent time spent index;
3. the average percent time spent per task in the TTM; and
4. the average percent members performing across TTM tasks.

The cumulative sum of percent time spent per module (TTM) was the only index not already computed in ASCII CODAP. Through coordination with AFHRL's Manpower and Personnel Division, this function was added to ASCII-CODAP MODULE programs, thus avoiding any need to write TDS-specific software for this part of the FUS (Yadrick et al., 1988).

The identification of all training courses and job assignment flows within a specialty proved to be a greater problem. For example, not all courses, such as field training detachment (FTD) and mobile training team (MTT) courses, are identified in the typical occupational survey report (OSR). AFR 50-5 lists AFS-specific and aircraft-specific courses but does not provide detailed information on course content. Personnel data files (UAR) contain very limited training information but do indicate professional military education (PME) courses attended. They also give assignment histories, but these often use AFS title or generic skill-level names rather than job titles equivalent to OSR jobs. Training management systems contain training attendance and completion records for formal courses by individual, but not by AFS. Cross-KPATH analysis was attempted with three specialties where more than one OSR was available, in an attempt to develop a comprehensive picture of the dynamic flows of personnel through jobs. In the final analysis, no existing single source could provide all of the training data needed, no existing sources provided suitable information on job flows, and no promising method of coordinating different sources was available.

For TDS purposes, such information was synthesized from various sources to identify AFS courses, relevant generic FTDs, PME programs, and

specialized training mechanisms, such as the Educational Subject Block Indices (ESBIs) used in AFS 811XX. A Job and Training History survey was sent to a representative sample of experienced job incumbents (and a random sample of first-job personnel). Respondents identified their present and previous jobs, and listed dates of attendance for all training programs; they could also write in additional training programs. These data were sorted and processed to estimate rates of attendance for the relevant courses for each job, attrition rates, and assignment flows (average length of assignment, PME course attendance points, etc.). Such information was compared to the OSR or data from other objective sources.

Graphic flow patterns were developed (see Figure 1), as well as concise narrative summaries. Such displays were validated with AFS SMEs in subsequent interviews or meetings whenever possible.

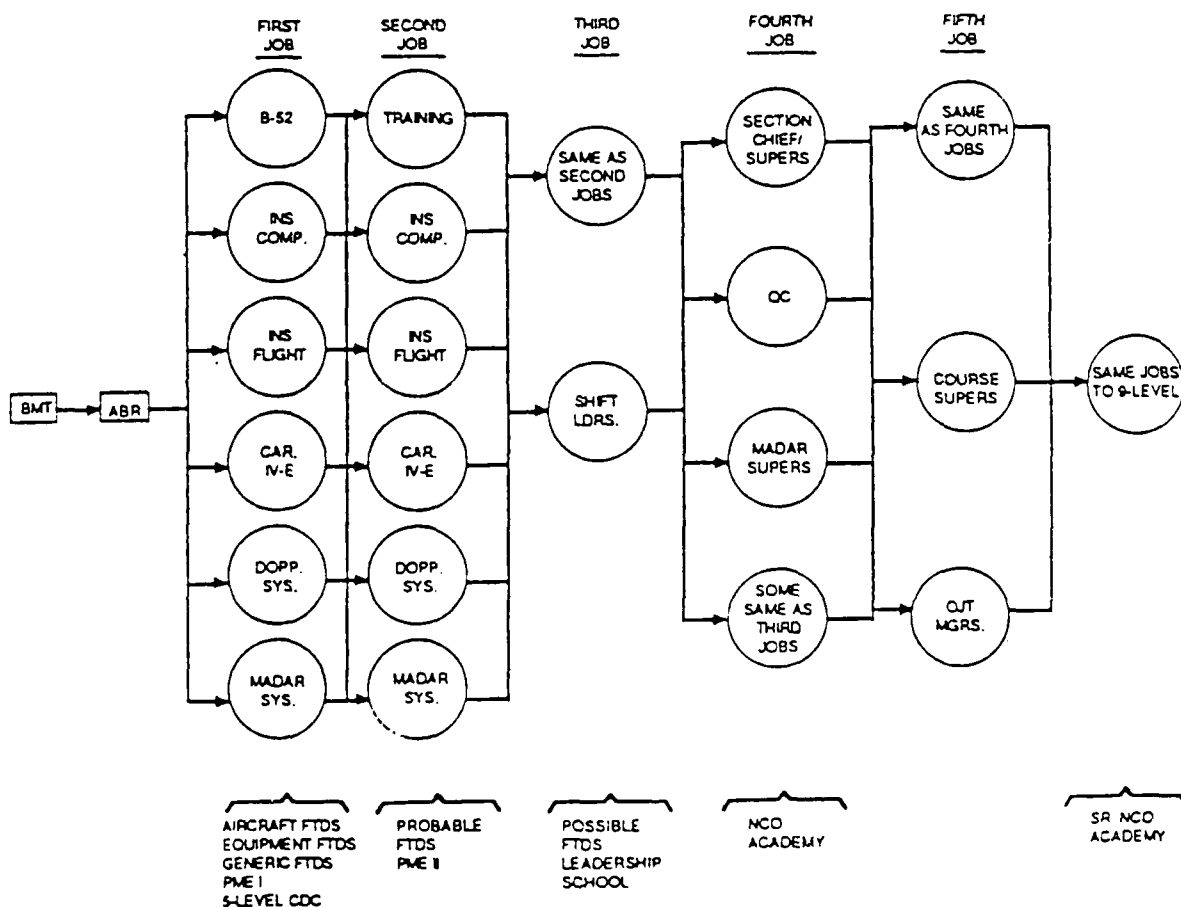


Figure 1. Sample U&T Pattern for AFS 328X4.

The graphic display provides a sense of the flow of individuals through training programs and jobs but does not lend itself to summarizing the various types of quantitative information involved, such as the number of individuals entering the field each year, or the various probabilities of reassignment among jobs, attending advanced technical training, or participating in PME courses. Such data, which are critical aspects of describing the specialty, are perhaps better conceptualized as a series of data matrices. Such data matrices summarize the probabilities of individuals being assigned to various jobs, of attending AFS and PME

training programs, and of exiting jobs for reassignment or to leave the AFS (or the Air Force). To adequately model the complex flows of personnel through AFS jobs and training programs, the TDS employs a dynamic simulation approach, which gives the analyst or researcher the capability to change specified values (i.e., number entering, assignment probabilities, attrition rates, etc.). Thus, the current U&T pattern can be modified to consider possible changes of the present approach to providing training or assigning personnel within the AFS; any such modification is considered an alternative U&T pattern.

Alternative U&T Pattern Component

The second FUS component develops alternative U&T patterns which managers and commanders might wish to consider and evaluate. Such alternatives could include patterns that minimize initial skills training (limit the number of first-term jobs), that represent present proposals for change of the specialty (e.g., the programmed expansion of Air Base Ground Defense training to all new Security and Law Enforcement personnel), that represent logical restructuring of work in an AFS, that involve reorganization of training for a specialty, or that might possibly result from expected procurement of new equipment or new procedures. TDS handles most proposed changes as alternative U&T patterns and builds displays and data files to simulate the consequences of these changes. Figure 2 displays an example of an alternative U&T pattern for the AFS represented in Figure 1.

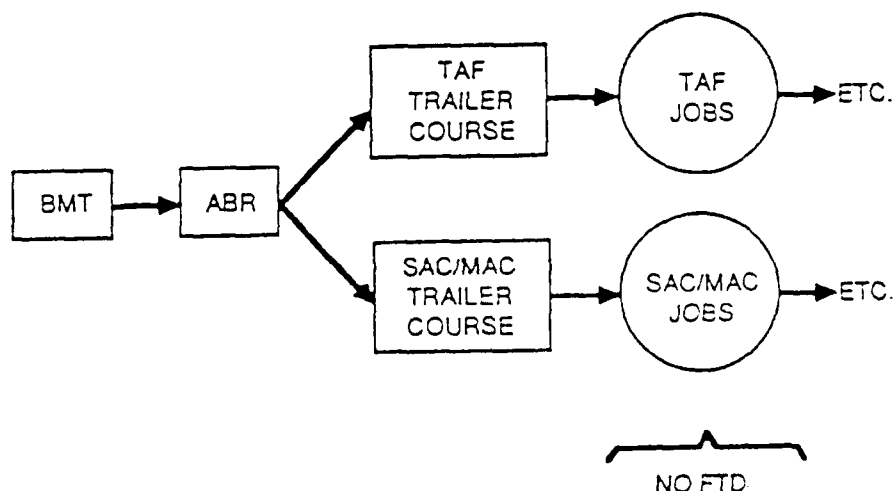


Figure 2. Example of a TDS Alternative U&T Pattern

Ideas for alternative patterns were developed from interviews with HQ USAF, major commands (MAJCOMs), and functional managers, supplemented by contacts with technical school instructors and other SMEs. Other alternatives were constructed based on rational job engineering or to meet some objective function (i.e., reduce the number of initial job possibilities, minimize initial skills training, etc.). In most cases, the potential alternative U&T model was created by modifying some major element of the current U&T data; for example, the simplest change is to increase or decrease the flow of personnel entering the specialty.

In all cases, the implications of change were evaluated in terms of impact on flow of assignments, school attendance rates, etc. The structure of the FUS model facilitates this process since it forces consideration of how changes can be implemented as modifications to the data.

FUS Flow Simulation Program

Computer programs were written for TDS to facilitate analysis of current and alternative U&T patterns, and to permit calculation of total AFS training requirements for each pattern. The main program is a dynamic simulation system which processes descriptive data files in such a way as to show total numbers of personnel flowing through jobs and training programs over an extended period of time. Data can be examined by any specified period of time. Training flows are generally expressed as annual rates; assignment probabilities are more realistically portrayed for 2- or 3-year intervals equivalent to the typical AFS job assignment.

New simulation software was developed which would better meet the needs of the TDS. The resulting program is titled UTPSIM, the U&T Pattern Simulation Program. It generates a flow pattern of entities (hypothetical individuals) moving from the initial training course or courses through first jobs (and associated training) to advanced courses or new jobs over a full career. It also accounts for career field attrition (both cross-training to other specialties and leaving the Air Force), FTD and PME requirements, and other AFS-specific factors.

The output of the UTPSIM program is an Output Entity History File (OUTEHF) which contains the job and formal training history of each individual (entity). For any given set of input data (current or alternative U&T patterns, or variations of input parameters, such as annual manpower input), the dynamic simulation calculates how individual entities enter the system, move through training and job states, and exit the AFS (or the Air Force). These data are reported as month-to-month summary flow statistics. Another program, the Training Proficiency (TRNPRF) program, then computes the total amount of OJT needed in the specialty for all individuals to achieve required proficiency on the tasks of the TIMs involved in their jobs.

The TRNPRF program also requires an Allocation Curve file, which contains parameters derived in the TCS to reflect how training is to be allocated to various training settings, and a TIM-Course file containing data collected via a survey of TTC and FTD instructors as well as OJT trainers in representative field units. These data characterize courses in terms of the number of hours of classroom time, hands-on training, and self-study time for relevant TIMs. Data for all courses of the specialty are arrayed in a matrix of courses by TIMs; cell entries are the hours of instruction (classroom plus hands-on training) provided in the course. Details of the UTPSIM and TRNPRF programs are available in other TDS documents (e.g., Vaughan, Mitchell, Marshall, Feldsott, & Rueter, 1988).

The FUS and its simulation programs play a key role in the overall TDS in terms of developing a quantitative data base which characterizes a particular U&T pattern. The resulting models (current and alternative U&T patterns) can thus serve as a foundation for estimating training costs and capacities for the specialty in the Resource/Cost Subsystem.

References

- Air Force Regulation 50-5 (1986, 1 June). USAF formal schools. Washington, DC: Headquarters, United States Air Force.
- Archer, W.B. (1966). Computation of group job descriptions from occupational survey data (PRL-TR-66-12, AD-653 543). Lackland AFB, TX: Personnel Research Laboratory.
- Christal, R.E. (1974). The United States Air Force occupational research project (AFHRL-TR-73-75, AD-774 574). Lackland AFB, TX: Occupational Research Division, Air Force Human Resources Laboratory.
- Driskill, W.E., & Mitchell, J.L. (1979, October). Variance within occupations; jobs analysis versus occupational analysis. Proceedings of the 21st Annual Conference of the Military Testing Association. San Diego, CA: Navy Personnel Research and Development Center.
- Morsh, J.E. (1964). Job analysis in the United States Air Force. Personnel Psychology, 17, 7-17.
- Phalen, W.J., Staley, M.R., & Mitchell, J.L. (1987). New ASCII CODAP programs and products for interpreting hierarchical and non-hierarchical task clusters. Proceedings of the Sixth International Occupational Analysts' Workshop. San Antonio, TX: USAF Occupational Measurement Center.
- Shartle, C.L. (1959). Occupational information (3d ed.). Englewood Cliffs, NJ: Prentice Hall, Inc.
- Vaughan, D.S. (1978, October-November). Two applications of occupational survey data in making training decisions. Proceedings of the 20th Annual Conference of the Military Testing Association (Vol. 1). Oklahoma City, OK: U.S. Coast Guard Institute (214-215).
- Vaughan, D.S., Mitchell, J.L., Marshall, G.A., Feldsott, S., & Rueter, F.H. (1988, August). Training decisions system procedural guide: TDS user instructions. (draft Technical Report, CDRL 25). Brooks AFB, TX: Prepared for the Training Systems Division, Air Force Human Resources Laboratory.
- Yadrick, R.M., Knight, J.R., Mitchell, J.L., Vaughan, D.S., & Perrin, B.M. (1988, July). Training decisions system: Development of the field utilization subsystem (AFHRL-TR-88-7). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.

RESOURCE/COST SUBSYSTEM:
ESTIMATING TRAINING CAPACITIES AND COSTS

F. H. Rueter and Steve Feldsott
CONSAD Research Corporation

The RCS was developed to provide TDS three distinct, yet interrelated, analytic capabilities:

1. determination of the types and amounts of resources required to provide training on each TIM in each training setting, and estimation of the amounts of those resources available for use in providing training at various sites;
2. assessment of the capacities of sites to accommodate different volumes of training on different combinations of TIMs in different training states, where a training state consists of a set of specific amounts of training conducted on specific TIMs in particular training settings; and
3. estimation of the variable costs that must be incurred in providing training on each TIM in each training setting, and in providing particular volumes of training in specific training states.

To accomplish these objectives, the RCS is structured into three analytic components: the Resource Requirements Component (RRC), the Training Capacity Component (TCC), and the Cost Estimation Component (CEC). These components use input files from the TCS and FUS; compile resource requirements, availability and cost factor data; and interact with one another to generate capacity and cost estimates for specific utilization and training (U&T) patterns (see Figure 1; see also Rueter, Vaughan, & Feldsott, 1988).

Resource Requirements Component (RRC)

The RRC performs five data development functions. For individual specialties, it (a) determines the specific types of resources required to perform training on each TIM of the specialty in each training setting, (b) estimates the quantity of each identified type of resource required for training each TIM in each setting, (c) produces compilations of those estimates classified on the basis of the ways in which the corresponding types of resources affect variable training costs and training capacities, (d) estimates the quantities of those resources available for the provision of training at various actual sites, and (e) delineates an appropriate set of representative sites for the particular specialty.

Inputs to this component include: TIM definitions and amounts of time allocated for training the various TIMs in different training settings (from the TCS), and preliminary lists of resources required for training each TIM in each setting (developed from available Air Force data and refined in collaboration with subject-matter experts). Data indicating

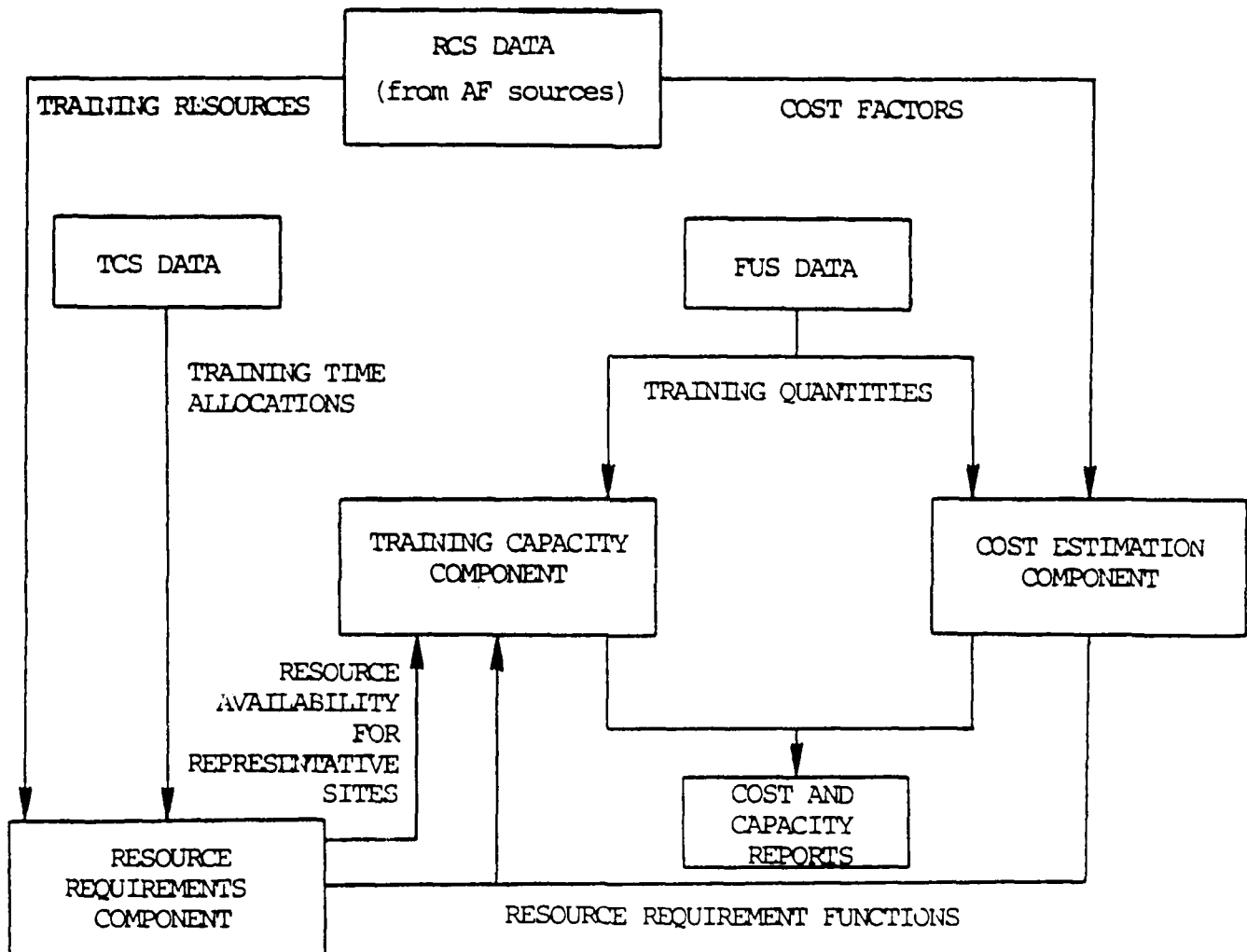


Figure 1. The TDS Resource/Cost Subsystem (RCS).

the amounts of those resources required for the provision of training in each setting were then collected, via a Training Resources Requirements Questionnaire administered to personnel in various jobs throughout the specialty. Data indicating the amounts of the resources typically available for use in training were also collected, via a Training Resources Availability questionnaire administered to senior enlisted personnel in each Air Force unit to which personnel of the AFS are assigned. Based on these inputs, the basic data used in estimating training capacities and costs within the other two RCS components are developed.

Representative sites (TTC courses, FTD courses, operational units) are identified to account for important variations in training loads, missions, resource availability, and other factors (Rueter, Vaughan, & Feldsott, 1987; 1988). The use of representative sites permits more economical collection of cost data and simplifies comparisons of resource availability and resource requirements. Each operational unit with personnel in the AFS is, however, uniquely associated with a particular representative site, thereby allowing suitable accounting for locational variations in travel and temporary duty (TDY) costs.

Training Capacity Component (TCC)

The Training Capacity Component evaluates the capacities of various representative sites to provide training in appropriate settings on different combinations of TIMs and in training volumes that are compatible with the U&T patterns identified in the FUS. Inputs to this component consist of the following: TIM combinations and training volumes for the various U&T patterns (from the FUS), predicted amounts of specific resources required for the provision of training on each TIM in each training setting (in the form of regression equations from the RRC), and availabilities of those resources for providing training at each representative site. Resource availability data are collected in a Training Resources Availability questionnaire administered at TTCs, FTDs, and operational field units.

An analyst, using analytic procedures contained in the TTC and IOS (Rueter, Bell, & Malloy, 1980), develops estimates of the capacity of each representative site to accommodate various combinations of TIMs and training loads, and identifies any limitations in the availability of specific types of resources or personnel that constrain representative sites from accommodating particular U&T patterns. When such constraints are encountered, they are displayed in the OJT Capacity Report for each site as "Trainees Unsupportable" (see Figure 2; note training deficit of 0.4959 trainees imposed by Resource 186). Training capacity evaluations use statistical training resource requirement functions and mathematical programming formulations to assess the feasibility of various training options.

Cost Estimation Component (CEC)

The CEC computes estimates of total annual variable costs for providing training of each TIM in each training setting (assuming all required resources are available in sufficient quantities), and then compiles the cost estimates in a form compatible with the estimates developed for training capacity. Inputs to this process include: estimated training

Representative Site: 5
 Training Capacity:
 - Upper Bound: 4.0360
 - Lower Bound: 4.0360
 Total Trainees Required: 4.5319

Resource ID	Amount Avail.	Amount Required	Avail/ Reqd Ratio	Maximum Trainees Sptable	Trainees Required	Trainees Not Sptable	Additional Amount Needed
39	21170.0	949.0	22.30	101.09	4.5319	0	0
103	5840.0	275.2	21.22	96.16	4.5319	0	0
134	7300.0	423.8	17.22	78.05	4.5319	0	0
150	2920.0	130.9	22.29	100.42	4.5051	0	0
172	5200.0	233.5	22.26	100.30	4.5051	0	0
177	1040.0	129.0	8.06	36.32	4.5051	0	0
186	730.0	819.7	0.89	4.03	4.5319	0.4959	89.70<<<
.							
.							
.							

Figure 2. Example of a Representative Site Training Capacity Report.

resource requirements (from the RRC), training states (i.e., amounts of time allocated to specific TIMs in specific settings) and associated training volumes compatible with various U&T patterns (from the FUS), and unit resource cost factors from external Air Force data sources (TDY costs, instructor salary levels, costs of training equipment and supplies, etc.). By applying the unit cost factors to the estimated training resource requirements for the specified training states and training volumes, this component estimates the variable costs of conducting training in each training setting and for each specified training volume in the corresponding training state (Vaughan et al., 1984).

Once these very complex basic data sets have been developed, they must be synthesized and processed as formatted reports in such a way as to be useful for Air Force decision makers. This is done by operation of the RCS components and data files as shown in Figure 3. The common starting point is the FUS Output File (from TRNPRF). Multiple processing is performed to generate training hours for trainees, trainers, and other resources, for both classroom and QJT requirements. It should be noted that the capacity and costing programs of the RCS operate in parallel, use some common data files as input, and use some unique files as well.

The major products of the RCS are data files and reports. The reports consist of capacity reports (see Figure 3) and cost reports. The cost estimates for QJT are reported by job and organizational level (unit, MAJCOM, total Air Force); the cost estimates for training in TTC and FTD courses indicate, for each course, the direct training costs (for trainees and trainers), travel costs, per diem costs, and cost per student week. Separate reports of training capacity and costs are generated for the current U&T pattern and each alternative U&T pattern to create multiple RCS output files. RCS data files serve as the basis for comparing the impact of various suggested AFS changes and for generating reports to

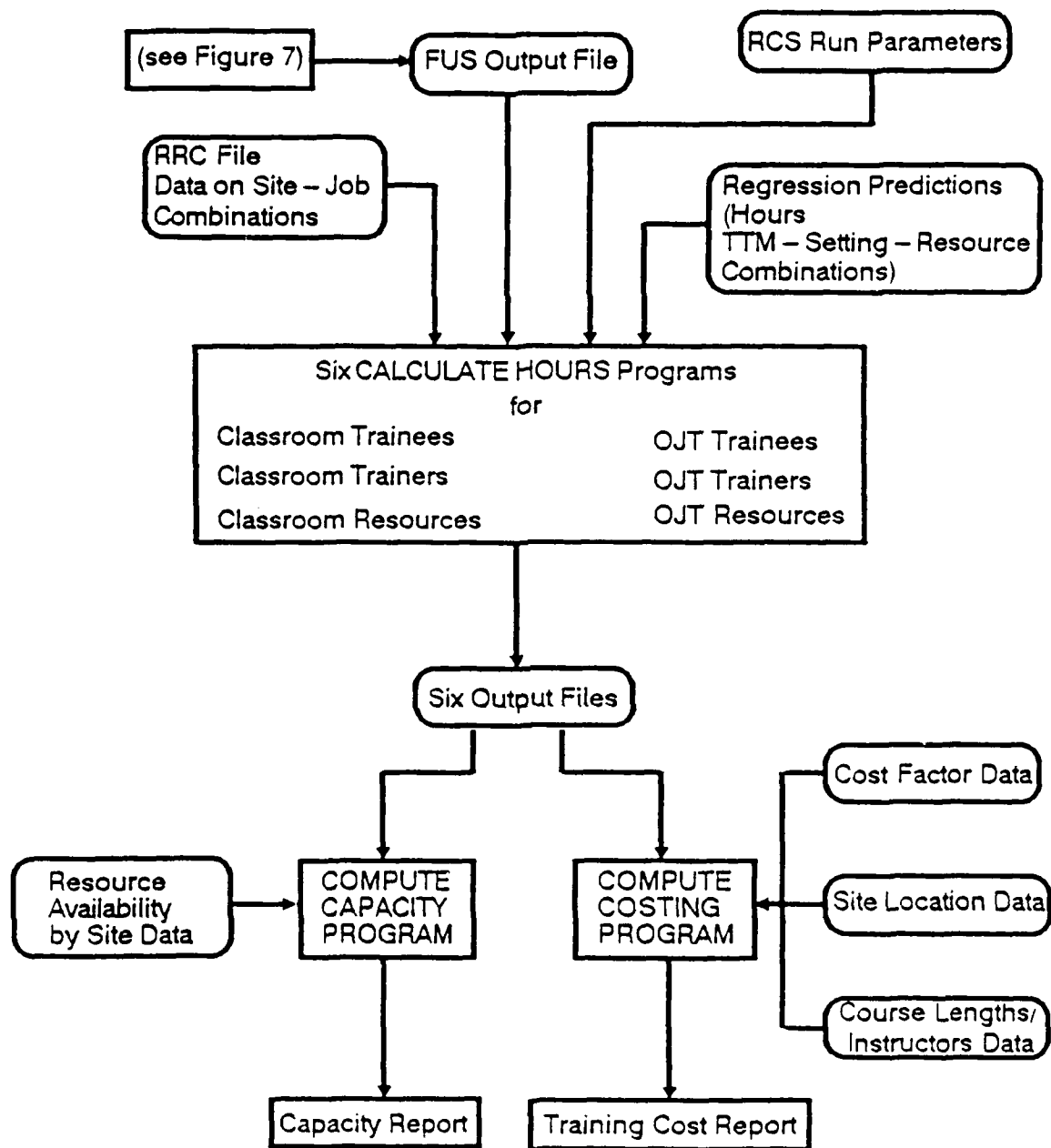


Figure 3. Interrelationships Among RCS Inputs, Components, and Products.

respond to managers' inquiries, through the operations of the Integration and Optimization Subsystem (IOS).

In the TDS development effort to date, the designed approach to the collection of resource information and calculation of unit level and total training costs for a specialty appeared to yield comprehensive and realistic results (Rueter, Vaughan, & Feldsott, 1988). Preliminary training sessions with ATC and AFHRL analysts and programmers were successful and feedback from trainees indicated they understood the RCS data and felt it should be extremely useful to Air Force decision makers. TDS training participants also provided very useful suggestions for improvements in product formats and the procedural guidance for operation of the system (Vaughan et al., 1988). Although there are a number of improvements which might be made in the RCS and other TDS subsystems (Vaughan et al., 1985), it appears to have succeeded in providing a methodology with the potential for quantifying the very elusive costs of on-the-job training as well as more formal training programs.

References

- Rueter, F.H., Bell, T.R., & Malloy, E.V. (1980, October). Capacity of Air Force operational units to conduct on-the-job training: Development of estimation methodology (AFHRL-TR-80-46; AD-A091 228). Lowry AFB, CO: Logistics and Technical Training Division, Air Force Human Resources Laboratory.
- Rueter, F.H., Vaughan, D.S., & Feldsott, S. (1987). The resource/cost subsystem (RCS) of the training decisions system (TDS): Project design (draft Technical Report, CDRL 20b). Brooks AFB, TX: Prepared for the Training Systems Division, Air Force Human Resources Laboratory.
- Rueter, F.H., Vaughan, D.S., & Feldsott, S. (1988). The resource/cost subsystem (RCS): Final administrative report (draft Technical Report, CDRL 22c). Brooks AFB, TX: Prepared for the Training Systems Division, Air Force Human Resources Laboratory.
- Vaughan, D.S., Mitchell, J.L., Marshall, G.A., Feldsott, S., & Rueter, F.H. (1988, August). Training decisions system procedural guide: TDS user instructions. (draft Technical Report, CDRL 25). Brooks AFB, TX: Prepared for the Training Systems Division, Air Force Human Resources Laboratory.
- Vaughan, D.S., Yadrick, R.M., Perrin, B.M., Cooley, P.C., Duntelman, G.H., Clark, B. L., & Rueter, F. H. (1984, August). Training decisions system: preliminary design (draft Technical Report, CDRL 21). Brooks AFB, TX: Prepared for the Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Vaughan, D. S., Yadrick R. M., Perrin, B. M., Mitchell, J. L., Sturdevant, W. S., Rueter, F. H., & Ward, Joe, Jr. (1985, September). Training decisions system transition plan. (draft Technical Report, CDRL 28). Brooks AFB, TX: Prepared for the Manpower and Personnel Division, Air Force Human Resources Laboratory.

INTERGRATION/OPTIMIZATION SUBSYSTEM:
AN INTEGRATED MODELING APPROACH

David. S. Vaughan
A. John Eschenbrenner
McDonnell Douglas Astronautics

The Integration and Optimization Subsystem (IOS) ties together the other three Training Decisions System (TDS) subsystems into one overall functional system (Collins et al., 1987; Vaughan et al., 1984). The interconnections of the TDS subsystems provide the capacity to optimize measures derived from one subsystem relative to constraints obtained from the others, and to simultaneously process data files derived from different subsystems. The IOS also provides the interface with users; that is, the subsystem receives all requests, calls appropriate data from the other subsystems or TDS files, and creates products to meet the users' needs. In all of its functions, the IOS governs the interaction among the TDS subsystems and various data sources, and the relationship of the system with various types of users.

Modeling Functions

The IOS processes information and data files from the TCS, FUS, and RCS to create various models of the AFS under consideration. The basic model for the AFS is the current U&T pattern; each alternative U&T pattern represents some change to the current U&T model. This approach provides maximum flexibility in the TDS, since an almost infinite number of possibilities can be considered.

The modeling functions of the IOS are not limited to examining the alternative U&T patterns formulated in the FUS (although AFS managers' preferences are collected only for these alternatives). Rather, IOS modeling provides the capability to change any input variable for any program, thus permitting examination of the impact of such change on the total system. For example, the simplest change would be to use the current FUS model and raise or lower the number of personnel entering the ABR course (modify the Trained Personnel Requirement or TPR). The system would then generate reports for comparison with the data from a baseline current U&T pattern run; differences in values (annual training costs, total AFS population in future years, etc.) would reflect the relative impact of the change. Another type of change would be to change the content of some course, such as the ABR, and then run the system to assess the changes in costs and total on-the-job training (OJT) requirements for the specialty.

Any major proposed change in a specialty should be dealt with as a formal alternative U&T pattern, so that possible consequences can be examined in some detail and data collected as to Air Force managers' preferences among a set of alternatives. Some changes, such as a merger of several AFSs at the technician level (as proposed in RIVET WORKFORCE), may exceed present system capacities (unless a new OSR is accomplished using a redeveloped task list covering all specialties involved, or needed data are estimated in some other way).

Most of the possible changes which might be considered for an AFS can be modeled by changing input parameters, course content, job content, or career patterns within the TDS. Such changes are processed as modeling runs with specified values of selected variables. Analysis then focuses on how such changes impact training capacities and total training costs. The training capacity reports generated in this process will also highlight any constraints on training capability in terms of training equipment, instructor availability, or other significant problems.

The example used in the earlier overview of the TDS provided some perspective on how changes in training one piece of test equipment (the Weapons Release Control System Analyzer in AFS 328X4, Radar and Inertial Navigation Systems Maintenance) might be examined and evaluated. The result in that example was to show the economy and practicality of moving the WRCS Analyzer training from an OJT setting to the basic resident course at the Technical Training Center.

Another example might be to eliminate entry-level formal training altogether; here our example specialty is the Aircraft Environmental System, AFS 423X1. Results of the current U&T model for this AFS and an alternative model with no entry-level course are shown in Figure 1 below.

	<u>CURRENT</u>	<u>NO ABR COURSE</u>	<u>DIFFERENCE</u>
ENTRY-LEVEL COURSE	\$941,517	\$ - 0 -	- \$941.517
OJT COSTS	\$530,189	\$808,417	\$278,228
OJT HOURS	55,662 HRS	85,038 HRS	29,416 HRS
OJT CAPACITY	NOT EXCEEDED	<u>EXCEEDED</u>	

Figure 1. Results of Eliminating AFS 423X1 ABR Course

As can be seen from these results, eliminating the basic resident course for the AFS does save money, but also exceeds the capacity of field units to provide OJT. Thus, this is not an effective solution unless additional resources (manpower and/or equipment) can be provided. The OJT Capacity report for the NO ABR Course model would identify those resources which were at issue, as a way of assisting managers to determine if this change was even possible.

Another kind of AFS change might be to alter the patterns of training and jobs in order to maximize the crossflow among aircraft systems. This type of crossflow is often suggested as a way to develop a versatile work force. In the Inertial and Radar Navigations Systems specialty (AFS 328X4), some functional managers suggested this type of crossflow could be brought about by making all first assignments to Tactical Air Forces jobs (small aircraft) with second jobs being with Strategic or Airlift systems (big aircraft). A model representing such an alternative was developed and is compared with the present U&T pattern in Figure 2.

	<u>CURRENT</u>	<u>VERSATILE WORK FORCE</u>	<u>DIFFERENCE</u>
FORMAL TRAINING	\$2,724,296	\$2,608,397	- \$115,899
OJT COST	\$5,096,500	\$5,009,593	- \$ 86,907
OJT CAPACITY	NOT EXCEEDED	NOT EXCEEDED	

Figure 2. Versatile Work Force (TAF --> SAC/MAC Assignments).

This alternative job flow does create a more versatile work force in terms of forcing succeeding assignments with different aircraft systems. At the same time, this solution will save money (which was not an expected result) and does not exceed the OJT capacity of field units. Thus, this appears to be a very viable solution which could be seriously staffed as a realistic proposal for change.

These examples demonstrate that the TDS can model both the training programs and the job structure of Air Force specialties, including possible changes in career paths available to AFS incumbents (Mitchell, Vaughan, Yadrick, Collins, & Hernandez, 1988; Vaughan, Mitchell, Marshall, Feldsott, & Rueter, 1988). The last example demonstrated the TDS capability to model job-to-job flows and assess the possible impact in terms of job task content and training costs and capacities.

The point here is that once a problem and potential solutions have been identified, the IOS modeling capability can be used to translate the potential solutions into modifications of the AFS data base (the current U&T files). The TDS software is then employed to generate products for each potential solution and results can be compared to baseline data (current U&T products) by Air Force decision makers (see Figure 3).

Optimization Functions

Given the almost limitless number of possible changes which might be studied, a TDS analyst or functional user may wish to take another approach to assessing AFS training changes. This approach would make use of the IOS optimization software. The analyst or user would specify an objective function or goal (such as minimization of OJT cost or total training costs, or maximization of the amount of available equipment, etc.), run the optimization program, and examine the effects on the AFS if the objective function is maximized or minimized. The analyst can ask "What if" questions; for example:

What is the impact on total training costs if we minimize initial resident course instruction?

What happens to AFS jobs (proficiency), if we maximize FTD training and minimize OJT?

What is the impact on AFS proficiency, acquisition and training costs, if we have a 10% cut in new recruits entering training?

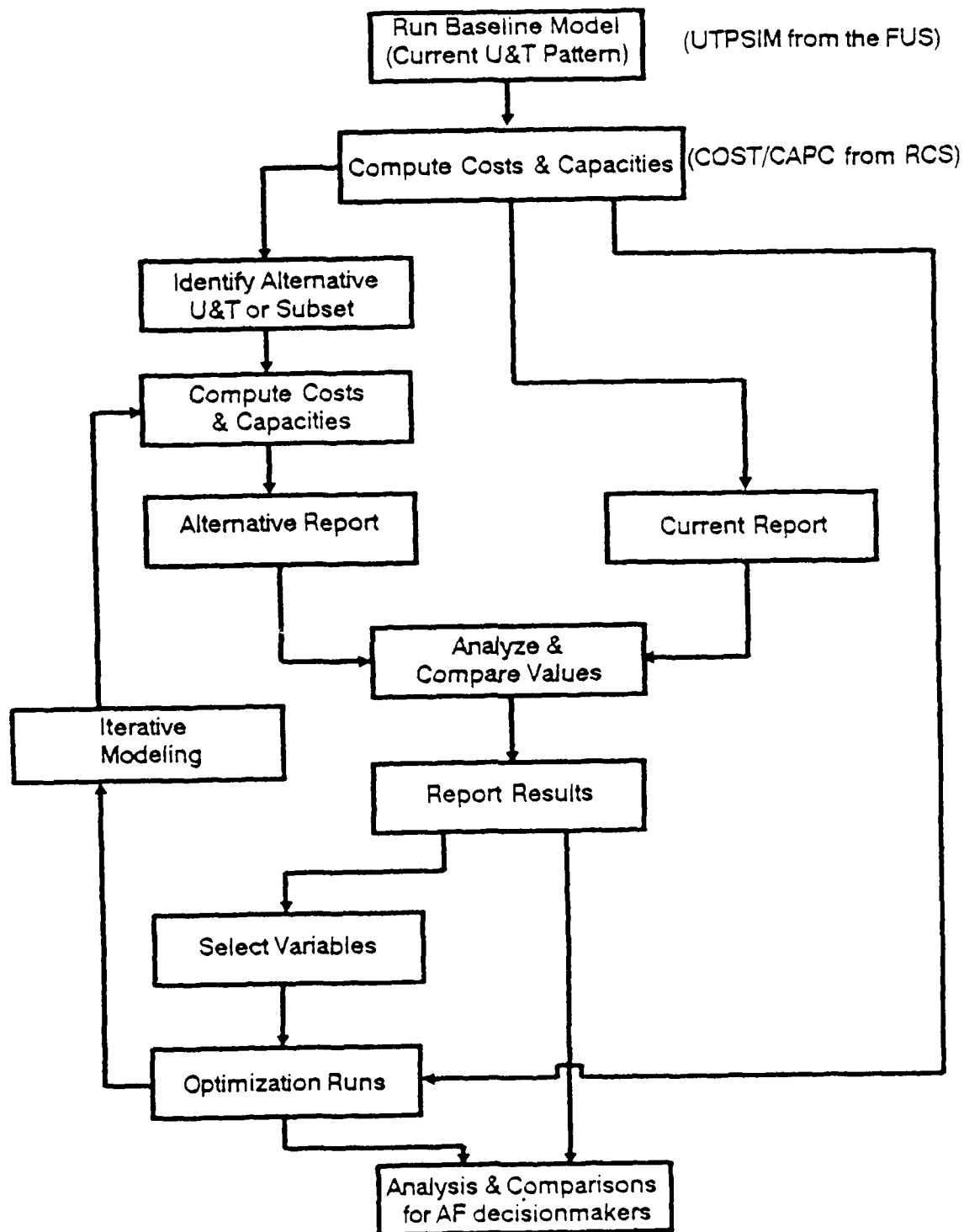


Figure 3. Modeling and Optimization Operations of the IOS.

Some potential optimization problems will become visible during modeling runs of the AFS as training constraints are identified, or possible new models may be suggested by initial optimization runs (see Figure 3). Other possible optimizations will be suggested by general Air Force trends, such as budget cuts or changing operational priorities. In some cases, these could be complex problems with several constraints and multiple values to be optimized.

The approach taken in employing optimization algorithms to solve maximization or minimization problems in the TDS is to employ modular data bases and seek solutions at the lowest possible level (Vaughan, Yadrick, Perrin, Cooley, Dunteman, Clark, & Rueter, 1984; Vaughan, Mitchell, Yadrick, Perrin, Knight, Eschenbrenner, Rueter, & Feldsott, 1988). This isolates solutions to only the area of interest and has considerable efficiency in terms of saving computer time. Only the largest optimization problems, such as minimizing total AFS training costs, would require employment of the entire TDS data base. More limited problems can thus be dealt with by limited program runs.

TDS Test & Evaluation

The research and development of the TDS has been successfully concluded, and the system has undergone preliminary Test and Evaluation (T&E) by AFHRL and ATC analysts and programmers, who were provided a TDS Users' Guide (Vaughan, Mitchell, Marshall, Feldsott, & Rueter, 1988). This T&E resulted in identification of some areas where improvements can be made, particularly in terms of formatting of products to be more useable by Air Force decision makers. Some suggestions were also made for additional software development to further enhance the capabilities of the TDS. Overall, the system was well accepted and the participants in the T&E expressed a very positive evaluation of the potential for the system to make significant improvements in Air Force training decision making. In addition, a TDS Transition Plan was developed previously to guide the operational implementation of the system (Vaughan et al., 1985).

Conclusions

The TDS approach to modeling occupational areas is a powerful tool for strategic human resources planning. Though the system is designed primarily for use by training decision makers, it is impossible to separate that use from the broader arena of manpower and personnel policies and programs. The integrated approach which has been used in the design of the TDS has very great potential for not only enhancing training decisions but also for promoting the eventual integration of total manpower, personnel, and training management.

References

- Collins, D.L., Hernandez, J.M., Ruck, H.W., Vaughan, D.S., Mitchell, J.L., & Rueter, F.H. (1987, August). Training decisions system: Overview, design, and data requirements (AFHRL-TP-87-25, AD-A183 978). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Mitchell, J.L., Vaughan, D.S., Yadrick, R.M., Collins, D.L., & Hernandez, J.M. (1988). The Air Force training decisions system: Modeling job and training flows (AFHRL-TP-88-12). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Vaughan, D.S., Mitchell, J.L., Marshall, G.A., Feldsott, S., & Rueter, F.H. (1988, August). Training decisions system procedural guide: TDS user instructions. (draft Technical Report, CDRL 25). Brooks AFB, TX: Prepared for the Training Systems Division, Air Force Human Resources Laboratory.
- Vaughan, D.S., Mitchell, J.L., Yadrick, R.M., Perrin, B.M., Knight, J.R., Eschenbrenner, A.J., Rueter, F.H., & Feldsott, S. (1988, October). Research and Development of the Training Decisions System (draft AFHRL-TR-88- 50). Brooks AFB, TX: Training Systems Division, Air Force Human Resources Laboratory.
- Vaughan, D. S., Yadrick, R. M., Perrin, B. M., Cooley, P. C., Dunteman, G. H., Clark, B. L., & Rueter, F. H. (1984, August). Training decisions system preliminary design (draft Technical Report, CDRL 21). Brooks AFB, TX: Prepared for the Manpower and Personnel Division, Air Force Human Resources Laboratory.
- Vaughan, D. S., Yadrick, R. M., Perrin, B. M., Mitchell, J. L., Sturdevant, W. S., Rueter, F. H., & Ward, Joe, Jr. (1985, September). Training decisions system transition plan. (draft Technical Report, CDRL 28). Brooks AFB, TX: Prepared for the Manpower and Personnel Division, Air Force Human Resources Laboratory.

DISCUSSION OF THE TDS PROJECT

Hendrick W. Ruck, Technical Advisor
Training Systems Division
Air Force Human Resources Laboratory

The development of the Training Decisions System (TDS) reported in this symposium represents the culmination of a line of research which began in the mid-1970's, and which has progressed step-by-step to the present evolved, computer-based, decision support system. The TDS is designed primarily to provide Air Force decision makers with information in a form which will assist them in making critical decisions involving training. The operational concept involved is to be able to show decision makers some of the possible consequences of their decisions in specific quantitative terms, something which has never been done before. The very complex TDS challenge involved designing new methods to collect, aggregate, predict, and display Air Force specialty (AFS) job and training data, including modeling alternative specialty patterns and estimation of capacities and costs for formal courses as well as on-the-job training (OJT).

This demonstration project has exceeded expectation; even those areas initially considered high risk (such as OJT costing) have been handled well in the present systems design. In addition to completing the TDS development, the research team at our request devised a PC-based TDS demonstrator to simulate operation of the system, which was then used to illustrate the concept for potential users. Further, a TDS transition plan was constructed, which provides insights into how to operationally implement the TDS; potential users and operators were identified as well as manpower and computer resources needed to support the system. Initial test and evaluation of the TDS demonstrated many of its capabilities; training was provided to AFHRL and Air Training Command (ATC) programmers and analysts who conducted further evaluations and suggested additional possible uses and capabilities which might be developed. A final report of the TDS research and development project has been completed and reviewed; it includes a proposed Air Force regulation for formal implementation of the system and a parallel research program. The TDS final report will be available as AFHRL-TR-88-50 early next year. I have been out briefing senior Air Force managers on the system over the last few months and their reactions to the TDS are highly positive and enthusiastic.

Several other potential applications of the TDS are feasible and will be included in future research and development work. One involves multiple-specialty studies to consider major AFS mergers, much like the RIVET WORKFORCE initiative undertaken by HQ USAF in recent years. A multiple-AFS TDS study would be able to quantify the possible cost and capacity consequences of proposed mergers, including the impact of changes in OJT requirements. Another possible adaptation of TDS would be in the new weapons systems acquisition process, to examine possible MPT requirements and costs. The TDS will require some modification to meet this kind of need, but the changes are both feasible and practical.

Overall, AFHRL is very satisfied with the outcomes of the TDS project. Our point of contact for TDS is Mr. Winston R. Bennett, AFHRL/IDET, Brooks AFB, TX 78235-5001. (512) 536-3047.

TRAINING, RETENTION, AND SUSTAINMENT OF FOREIGN LANGUAGE
SKILLS IN THE ARMY

Zita M. Simutis, Chair
U.S. Army Research Institute
Alexandria, Virginia

Changes in military doctrine on operational readiness in response to world conditions have resulted in an increased demand for military personnel qualified to communicate in foreign languages. This panel focuses on the special challenges faced by military testers, trainers, and other researchers who make it possible to meet this increased demand. The panel begins with a description of a project that undertakes to examine the utility of two aptitude tests, the Defense Language Aptitude Battery (DLAB) and the Armed Services Vocational Aptitude Battery (ASVAB), used to select applicants for foreign language training. This is followed by an overview of an extensive language skill change project (LSCP) that is directed at determining the nature and extent of changes in foreign language skill levels among Army linguists after graduation from the Defense Language Institute Foreign Language Center (DLIFLC). The following paper examines the design and application to LSCP of an instrument for measuring the learning strategies and techniques that are useful for language retention and sustainment. A report on preliminary analyses extracted from the LSCP data that begin to delineate the influence of several factors (including aptitude, selection parameters and post-DLIFLC language training and use) on language skill change is presented in the succeeding paper. The final paper examines the role of new technologies to provide tools for two significant ends: understanding the cognitive processes that underlie language competence; and providing systems that can substantially aid and sustain language skills on the job.

A Preliminary Investigation of the Relationship Between the ASVAB and the DLAB¹

**Leonard A. White, Lawrence M. Hanser and Randolph K. Park²
U.S. Army Research Institute for the Behavioral and Social Sciences**

Introduction

Proficiency in a foreign language is required in several entry-level Military Occupational Specialties (MOS), and may be needed in other MOS for specific one-time assignments. To qualify for language training, applicants must exceed the minimum cut-off scores on the Defense Language Aptitude Battery (DLAB) and the Skilled Technical (ST) composite of the Armed Services Vocational Aptitude Battery (ASVAB). Training at the language schools is intensive, often requiring six hours per day in class, with homework assignments (Petersen & Al-Haik, 1976). Classes may last for over one year.

The current DLAB is a 119 item multiple choice test that requires 90 minutes to administer and use of an audio tape. Items on the DLAB are constructed to measure the most important abilities needed in learning to speak and understand a foreign language. These include the ability to process auditory phonetic material so that it can be recognized and recalled, grammatical sensitivity, and the capacity to learn a large number of new associations in a relatively short time (Carroll, 1972). To measure aptitude for foreign language learning the DLAB requires examinees to learn and use an artificial language. For example, examinees are asked to select a correct translation in an artificial language by applying a set of grammatical rules, to match pictures to phrases, and to recognize vowel "stress" patterns (Petersen & Al-Haik, 1977). The validity of DLAB as a predictor of success in language training has been established. Soldiers with high scores on the DLAB (as opposed to low) earn higher grades in language training ($r = .40$ uncorrected) and have lower rates of academic attrition (J. A. Lett, personal communication, August, 1987; Petersen & Al-Haik, 1977).

The current ASVAB is a conventional paper-and-pencil tests and is used to predict the "general trainability" of the examinees. Each ASVAB subtest consists of items with difficulty levels that span the range of abilities to be found in an applicant population. The 10 subtests on the current ASVAB (Forms 11, 12, and 13) are listed in Table 1. Eight of the tests are power tests and two tests (CS and NO) are speeded.

At the time of application for the Armed Services all Army recruits are tested on the ASVAB. Those interested in language training who fail to achieve the minimum score on the ASVAB are typically rejected from further consideration for language-dependent MOS. For those who meet the minimum

¹The views expressed in this paper are those of the authors and do not necessarily reflect the views of the U.S. Army Research Institute or the Department of the Army.

²Randolph Park is now at the American Association of Medical Colleges, Washington, DC.

Table 1

ASVAB Subtests

Subtest	Abbreviation	Number of Items	Time (Min.)
General Science	GS	25	11
Arithmetic Reasoning	AR	30	36
Word Knowledge	WK	35	11
Paragraph Comprehension	PC	15	13
Numerical Operations	NO	50	3
Coding Speed	CS	84	7
Auto and Shop Information	AS	25	11
Mathematics Knowledge	MK	25	24
Mechanical Comprehension	MC	25	19
Electronics Information	EI	20	9

ASVAB cut-off score, a final decision regarding eligibility for language training is based on the score received on the DLAB. A DLAB score of 85 or above is needed to qualify for training in low difficulty languages (e.g., Spanish). Those languages in the highest difficulty category (e.g., Chinese) require a minimum DLAB score of 100.

It has been suggested that this two-stage sequential decision plan would be more efficient if it included accepting as well as rejecting applicants for language school based on the ASVAB alone. The reasoning here is that those applicants with a high score (e.g., 120 or above) on ST or some other ASVAB composite are very likely to attain the minimum passing score on the DLAB. If there is a strong relationship between scores on the ASVAB and DLAB it would seem reasonable to accept applicants for language training based on ASVAB alone. This would reduce expensive DLAB testing time and may facilitate efforts to recruit candidates for language school.

To explore this possibility we examined relationships between scores on the ASVAB and DLAB in a sample of Army enlisted personnel. Of interest here, is to provide information on (a) the correlation between scores on ASVAB composites and the DLAB, and (b) the likelihood of attaining a passing score on the DLAB given certain cut-scores on the ASVAB.

Method**Subjects and Procedure**

The analyses reported in this paper are based on sample of 5,010 Army enlisted personnel who were administered the DLAB between September 1986 and January 1987. Scored DLAB answer sheets for Army examinees are sent to the Test Control Officer, at the Defense Language Institute Foreign Language Center (DLIFLC). The data set described in this paper was constructed by

recording the DLAB scores from examinee answer sheets available at DLIFLC and merging (by SSN) with ASVAB scores from enlistment records.³

In the enlisted soldier sample, 88% were male and 12% were female; 87% were Caucasian, 9% Black, 1.2% Asian, .5% American Indian, and 2.3% other. Also, of the 5010 soldiers in the sample, 80% entered the Army in FY86 or FY87, and the remaining 20% were FY84 or FY85 accessions.

Results and Discussion

The sample means and standard deviations on the current operational ASVAB composites and the DLAB are presented in Table 2. The AFQT score is reported as a percentile score. The mean AFQT score indicates that the average

Table 2

Means and Standard Deviations of Examinee Scores on the DLAB and ASVAB Composites

Test		Mean	S.D.
ASVAB Composites			
Armed Forces Qualification			
Test (AFQT)	VE + AR + NO/2	71.89	18.41
Clerical (CL)	AR + MK + VE	114.69	11.49
Combat (CO)	CS + AR + MC + AS	116.12	12.00
Electronics Rep'r (EL)	AR + MK + EI + GS	115.47	12.36
Field Artillery (FA)	AR + CS + MK + MC	116.09	12.20
General Maintenance (GM)	MK + EI + AS + GS	115.35	12.96
Mechanical Maintenance (MM)	NO + AS + MC + EI	115.71	12.21
Operators/Food (OF)	NO + AS + MC + VE	116.07	10.70
Surveillance (SC)	AR + AS + MC + VE	116.20	11.89
Skilled Technical (ST)	VE + MK + MC + GS	116.81	11.89
General Technical (GT)	VE + AR	114.27	10.58
Defense Language Aptitude Battery (DLAB)	All Items	77.5	19.8

Note. VE = PC + WK

³The authors wish to thank Frances Grafton, Faye Wyen, and John Lett and his staff at the Division of Evaluation and Research, DLIFLC, for their assistance in preparation of the data set.

examinee is in mental category II. The remaining ASVAB scores are standard score composites where the population mean is 100 and a standard deviation of 20. Note that the mean DLAB standard score ($\bar{X} = 77.5$) is below the passing score of 85.

Table 3 presents the uncorrected correlations between the DLAB scores and the ASVAB subtests and operational composites. Several results are noteworthy in Table 3. First, the correlations between scores on the ASVAB composites and DLAB are moderately high (range = .33 to .56). This suggests that the DLAB and ASVAB measure somewhat different aptitudes and a direct equating of DLAB with ASVAB may be problematic. Second, for ASVAB subtests, the highest correlations with the DLAB were obtained for the Arithmetic Reasoning subtest ($\bar{r} = .47$), and the Mathematics Knowledge subtest ($\bar{r} = .50$). Similarly, ASVAB composites with a high quantitative component showed the highest correlation

Table 3

Correlation Between ASVAB and DLAB Test Scores

ASVAB Subtest	Correlation with DLAB	ASVAB Composite	Correlation with DLAB
		GT	.52
GS	.36	CL	.56
AR	.47	CO	.43
WK	.40	EL	.51
PC	.36	FA	.54
NO	.21	GM	.41
CS	.25	MM	.33
AS	.11	OF	.38
MK	.50	SC	.42
MC	.33	ST	.51
EI	.27	AFQT	.53

with the DLAB. The correlation for the Clerical (CL) composite and DLAB was .56 and the correlation between the Field Artillery (FA) composite and DLAB was .54. These values should be compared against the operational composite ST which correlated with DLAB at .51.

The ST composite is used operationally to screen candidates from entering language-dependent MOS. Those applying for language training who meet the minimum ASVAB requirement are administered the DLAB. Table 4 presents relationships between examinees' ST score and the likelihood of attaining selected cut-off scores on the DLAB.

Overall, with current testing procedures, 34.5% of examinees achieved a score of 85 or more on the DLAB; 14% scored 100 or more on the DLAB. However, as can be seen in Table 4, examinees with a high ST score were more likely to

make a qualifying score on the DLAB. Of those scoring 130 or more on ST, 75.1% achieved a score of 85 or more on the DLAB. In contrast, one of ten applicants with a score of 104 and below on ST attained a minimum qualifying score on DLAB; only 1.3% scored 100 or more on the DLAB.

Table 4

Percentage of Examinees Exceeding Cut-off Scores on DLAB for Given Score Intervals on the ST Composite of the ASVAB

ST Score Interval	n	DLAB Score 85 and above	DLAB Score 100 and above
130 and above	650	75.1%	45.5%
125-129	865	54.2	24.1
120-124	1105	33.8	10.3
115-119	430	28.4	7.7
110-114	561	22.3	4.5
105-109	556	13.8	2.5
104 and below	843	9.9	1.3
Total Sample	5010	34.5	14.0

One purpose of this research was to determine if a cut-off score on ST (e.g., 120) could be located above which the probability of passing the DLAB was sufficiently high to support accepting applicants for language training on the basis of ST alone. To examine this question, the data summarized in Table 4 were used to compute the proportion of "hits" on the DLAB for given cut-offs on the ST composite. These relationships are shown in Table 5.

It is clear from Table 5 that setting a higher cut-off score on ST increased the percentage of examinees who make a qualifying score on DLAB. Of the examinees who scored 120 or above on ST, 51% scored 85 or above on the DLAB; only 24% scored 100 or more on the DLAB. Of these examinees who scored 125 or more on ST, 63% scored 85 or more on the DLAB; 33% scored 100 or above on the DLAB.

The CL composite had the highest correlation with the DLAB. Of those scoring 120 or above on CL, 59% scored 85 or more on the DLAB, and 29% achieved scores of 100 or more on the DLAB. Thus, as compared with ST, a cut-off score of 120 or more on CL yields a higher percentage of qualifying scores on the DLAB. However, in either case a substantial percentage of scores fell below the minimum passing DLAB score of 85.

Table 5

Percentage of Examinees Exceeding Cut-off Scores on DLAB for Given Cut-off Scores on the ST Composite of the ASVAB

ST Score	Cases	DLAB Score 85 and above	DLAB Score 100 and above
130 and above	650	75.1%	45.5%
125 and above	1515	63.2	33.3
120 and above	2620	50.7	23.6
115 and above	3050	47.6	21.4
110 and above	3611	43.7	18.7
105 and above	4167	39.7	16.6
100 and above	4631	36.7	15.0
Total Sample	5010	34.5	14.0

Summary

The results reported in this paper indicate that the efficiency of DLAB testing could be improved by raising from 95 to near 120 the minimum ST score (or possibly CL) required to take the DLAB. Raising the cut-off score on the ST composite would reduce the number of DLAB testing hours needed to identify a given number of applicants who make a qualifying score on the DLAB. The specific ASVAB cut-off score chosen should reflect consideration of the trade-offs between the cost of testing examinees who would otherwise not pass DLAB and minimizing the potential number of candidates being eliminated by the ST pre-screen process.

Relationships obtained between the ASVAB and DLAB were not sufficiently high to support selection of applicants for language training based on ASVAB alone, given that the DLAB is a predictor of success in language training. An important but as yet unanswered research question here, is the degree to which DLAB improves prediction of success in language training over ASVAB prediction. Reductions in DLAB testing may be justified if the DLAB offers little or no improvement over ASVAB in the prediction of success in foreign language learning. Research to investigate this issue is recommended.

References

- Carroll, J. B. (1982). Prediction of success in intensive foreign language training. In Robert Glaser (Ed.). Training research and education. Pittsburgh: University of Pittsburgh Press, 1962.
- Petersen, C. R. & Al-Haik, A. R. (1976). The development of the Defense Language Aptitude Battery (DLAB). Educational and Psychological Measurement, 36, 369-380.

The Language Skill Change Project (LSCP)

John A. Lett, Jr.¹

Defense Language Institute Foreign Language Center

Background

Each year significant amounts of fiscal and personnel resources are devoted to the acquisition of foreign language proficiency by members of the four uniformed services who are being trained to become military linguists. The vast majority of this language training is provided at the Defense Language Institute Foreign Language Center (DLIFLC) at the Presidio of Monterey, California, in courses ranging in length from twenty-five to forty-seven weeks. Because of the significant investment in the initial acquisition of foreign language skills, there has long been concern, both at DLIFLC and among user communities, regarding post-DLIFLC erosion of language skill. This concern led to the development of the Language Skill Change Project (LSCP), which is being conducted jointly by DLIFLC and the Army Research Institute (ARI) with the coordination and support of a Project Advisory Group (PAG) chaired by a representative from the US Army Intelligence Center and School (USAICS) at Fort Huachuca, Arizona.²

Although the principal objective of the LSCP is to assess the extent and nature of post-DLIFLC language skill change in Army Military Intelligence (MI) linguists, the study was designed to support an investigation of variables associated with initial foreign language acquisition as well. The purpose of this paper is to set the stage for a brief discussion of preliminary findings regarding the prediction of language training outcomes at DLIFLC on the basis of certain learner characteristics, and the possible relevance of these findings for the optimal selection and assignment of potential military linguists. The following topics are discussed: (1) The overall LSCP research design; (2) The predictor and criterion variables; (3) The analytical approach.

The LSCP Research Design³

Design. The LSCP is a longitudinal study in which each subject is to be tracked from entry into DLIFLC through advanced individual training (AIT) and post-AIT assignments for the next two years or until the end of the first enlistment period.

¹Dr. Lett is Chief of the DLIFLC Research Division. The views, opinions, and/or findings contained in this document are those of the author and should not be construed as the official position of DLIFLC or ARI nor as an official Department of the Army position, policy, or decision, unless so designated by other official documentation.

²Data collection, data base maintenance, and data analysis and reporting requirements are supported by Advanced Technology, Inc. under an OPM contract funded mainly by ARI and DLIFLC.

³More detailed descriptions of the information presented in this section and the next can be found in Kahn and Lett, 1985; Mutter, undated; and Bush, 1987.

Sample. LSCP subjects are defined as all US Army enlisted students entering DLIFLC from February 1986, through August 1987, who held or were being trained for the military intelligence (MI) military occupational specialties (MOS) of 97B, 97E, 98CL, and 98G,⁴ and who were assigned to study either Korean, Russian, German, or Spanish.⁵ This sample consists of 1903 individuals, distributed as shown in Table 1

TABLE 1

DISTRIBUTION OF LSCP SAMPLE BY LANGUAGE AND MOS

LANGUAGE	97B	97E	98C	98G	TOTAL
KOREAN	13	27	79	276	395
RUSSIAN	23	81	123	554	791
GERMAN	26	45	58	285	414
SPANISH	8	28	61	206	303
TOTALS	70	181	331	1321	1903

Measurement plan Data are gathered from LSCP students on at least six occasions: three times at DLIFLC, once at the end of AIT, and at each annual foreign language proficiency test thereafter. Data collection events and their purposes are as follows:

- Time 1: Prior to beginning their DLIFLC training, subjects completed an extensive array of questionnaires and inventories designed to measure variables thought to be relevant to the prediction of language learning outcomes.
- Time 2: Approximately twelve weeks into their language training, subjects completed an additional set of questionnaires similar to the first, but whose nature required that students have experienced a minimal amount of language instruction.
- Time 3: At the end of their DLIFLC language training, subjects completed a composite set of measures selected from among those previously administered at Times 1 and 2. At approximately this same time, LSCP subjects, along with their non-LSCP cohorts, were administered the appropriate Defense Language Proficiency Test (DLPT).⁶
- Time 4: At the end of AIT, LSCP students were administered another form of the DLPT-III, along with a questionnaire regarding the extent and nature of their language use and language maintenance activities since their departure from DLIFLC.⁷
- Times 5-6/7: At each of the annual language proficiency assessments mandated by AR 611-6, LSCP subjects and their immediate supervisors will be given a language use questionnaire similar to, but more extensive than, the AIT questionnaire.⁸

⁴MOS 97B and 97E duties require the use of all four language skills; 98CL and 98G duties involve primarily the use of listening and reading skills.

⁵A limited number of "bypasses" is also included at the request of one MI user community, these being Army MI personnel who did not acquire their foreign language skills at DLIFLC.

⁶All four LSCP languages are tested by the DLPT-III, not the earlier DLPT-I or DLPT-II.

⁷AIT data collection was scheduled to be completed in December 1988.

⁸Although the original time period between successive post AIT DLPT testing was set at six months, it was later lengthened to nine months for logistical reasons. Finally, with Foreign Language Proficiency Pay (FLPP) driving greater compliance with annual testing requirements, it was decided to utilize the latter for all post AIT DLPT measures.

Predictor and Criterion Variables

Criterion measures. As mentioned above, the DLPT-III constitutes the measure of language proficiency for this study.⁹ The DLPT-III measures language proficiency in three skill modalities, listening, reading, and speaking. Listening and reading are assessed via a computer-scorable multiple-choice test validated and normed in accordance with the official Interagency Language Roundtable (ILR) language proficiency level descriptions; speaking proficiency is assessed at DLIFLC in a face-to-face ILR oral interview; successive speaking tests are administered via tape recordings and test booklets; examinees' recorded responses are returned to DLIFLC for rating by certified testers. All DLPT-III materials are controlled test items.

Predictor variables. Predictor variables include both data routinely available in soldiers' official records and data gathered specifically for this study. They are described here in an approximate order of accessibility and increasing collection cost.

1. **General Ability.** The General Technical (GT) composite of the Armed Services Vocational Aptitude Battery (ASVAB) was selected by virtue of its availability and general nature, the GT being a composite of verbal subtests (word knowledge and paragraph comprehension) and arithmetic reasoning.
2. **Language Learning Aptitude.** The Defense Language Aptitude Battery (DLAB), required by all services as part of the selection process for future MI trainees, was used.
3. **Available Demographics.** Normal DLIFLC registration procedures generate a student record containing information thought to be relevant to language learning success. Variables utilized include sex, level of education, and age.¹⁰
4. **Handedness.** Because brain hemisphericity is thought by some researchers to be related to language learning, handedness was included in the special LSCP Language Background Questionnaire (LBQ) created for this study.
5. **Prior FL Experience and Proficiency.** The LBQ was also used to collect information regarding prior foreign language training experience and self-reported prior foreign language proficiency.
6. **Attitude/Motivation at Start and at Twelve Weeks.** The work of Robert Gardner over the last three decades (e.g., Gardner & Lambert, 1959; Gardner, et al., 1983) provided the nucleus for the measures of attitude and motivation used in the LSCP. With Gardner's assistance, many scale items were modified slightly for use in DLIFLC's intensive military training context. In addition, new Gardner-inspired scales measuring various aspects of instrumental motivation were created by DLIFLC and ARI researchers (Lett and Ekstrom) for use in this study. In addition to the Gardner Battery, Hiller's Personal Outlook Inventory (Kirby & Hiller, 1973) was used at the start as a measure of general intellectual self-confidence.
7. **Learning Strategies.** Under an ARI OPM contract, a Strategy Inventory for Language Learning (SILL) was created by Rebecca Oxford for use in this study (Oxford, 1986). The SILL was used at the 12-week point and at the end of DLIFLC training, and shortened forms are being used at AIT and in post AIT administrations.

⁹An additional, dichotomous, measure of language training success was constructed to measure the prediction of academic attrition.

¹⁰For the analyses reported by O'Mara in this volume, age is represented by the surrogate variable, military pay grade.

8. **Personality and Cognitive Style.** Measures of empathy, ambiguity tolerance, field independence, and extraversion were included in the Time 1 battery.

9. **Other Ability.** As a matter of convenience, three additional measures of ability were included in the Time 1 battery, i.e., the Watson-Glaser Critical Thinking Inventory (CTI), Flanagan's Industrial Test of Memory, and Flanagan's Industrial Test of Expression. Although these measures have no known history of use in the prediction of foreign language learning, and were originally included to support an ancillary study, they have turned out to be rather interesting components of the regression equations examined to date. Thus, they are reported as part of this data set.

The Analytical Approach

Although the data collected in this study will lend themselves to the investigation of various explanatory models of classroom-based language learning, for the present purpose we examined them from a predictive perspective in order to explore their potential for enhanced selection and assignment procedures for military linguists. Analysis procedures included data reduction via factor analysis and principal components analysis, followed by multiple regression analyses using a forward progression, forced order of entry approach. The results of these preliminary analyses are presented in O'Mara's paper in this volume. Subsequent analyses of the present data set will address explanatory models; analyses of AIT and field-based data will address the question of skill change over time and variables related thereto--including initial proficiency levels attained at DLIFLC.

References

Bush, B. The Language Skill Change Project (LSCP): Background, Procedures, and Preliminary Findings (1987). Research Report 1464, Army Research Institute Field Unit at Presidio of Monterey, CA.

Gardner, R.C., Lalonde, R.N., & Pierson, R. (1983). The socioeducational model of second language acquisition: An investigation using causal modeling. Journal of Language and Social Psychology 2(1), 1-15.

Gardner, R.C., & Lambert, W.E. (1959). Motivational variables in second language acquisition. Canadian Journal of Psychology, 13, 266-72.

Kahn, O., & Lett, J.A., Jr. (1985). Proposed research and measurement plan for Language Skill Change Study, Phase I. (Available from Research Division, DLIFLC, Presidio of Monterey, CA 93944-5006.)

Kirby, E.A., & Hiller, J.H. Comparative validation of a direct and an indirect measure of academic self-confidence. Paper presented at the meeting of the American Educational Research Association, New Orleans, February, 1973.

Mutter, S. (undated). Individual differences in second language acquisition and loss. Unpublished manuscript.

Oxford, R.L. (1986). Development and psychometric testing of the Strategy Inventory for Language Learning (SILL): Appendix (Research Note 86-92). Alexandria, VA: U.S. Army Research Institute, Department of the Army. ADA 175452 Nov 153pp

Development and Evolution of the Strategy Inventory for Language Learning

Rebecca L. Oxford, Ph.D.

Annenberg/CPB Project

This brief paper describes the development and evolution of the Strategy Inventory for Language Learning (SILL). The SILL was first developed in 1985 by the author of this paper for use in the Language Skill Change Project (LSCP), the purpose of which is to determine the variables responsible for changes in military intelligence careerists' foreign language skills after formal language training is completed. The LSCP is a joint effort of the Army Research Institute (ARI) and the Defense Language Institute (DLI).

Learning strategies, i.e., steps taken by the learner which are intended to facilitate the acquisition, retention, and retrieval of new information, may be an important factor in determining what is learned in the first place and what is eventually lost or maintained after the end of language training. Therefore, ARI and DLI decided it was important to include learning strategies among the variables to be studied in the LSCP. Since there was no language learning strategy instrument which had been demonstrated to be reliable, valid, and useful for the given purpose, ARI contracted to have such an instrument developed specifically for use in the LSCP. The result was the SILL. Since that time, because of the strong need for such an instrument worldwide, the SILL has been used in exploratory studies and for counseling purposes in a number of settings beyond the LSCP; for instance, it has been used in several dissertations and by organizations such as the Peace Corps, the Foreign Service Institute, the Center for Applied Linguistics, Purdue University, the University of Hawaii, the University of Maryland, and universities and schools in Indonesia, Spain, Guatemala, and Taiwan. It is being translated into several other languages, and a shortened and simplified form has been created for use by non-native speakers of English enrolled in programs in English as a second or foreign language.

Initial Steps

The author conducted an extensive research review on second and foreign language learning strategies. Using the research

review, the author then developed a comprehensive taxonomy of language learning strategies and later expanded the taxonomy to show how each strategy could be related to each of the four language skills (reading, listening, speaking, and writing). The SILL items were based on the taxonomy. A 23-person clinical trial was held to check for readability, social desirability response bias, clarity of items, usability of format, and other details.

Initial Reliability and Validity Assessment and Factor Analysis

In late 1985 a 483-person field test (Study 1) was conducted at the Defense Language Institute to assess the reliability and validity of the SILL, to check again for social desirability response bias, and to obtain an initial factor analysis of the instrument.

Results indicated that the SILL had very high reliability and validity coefficients. The internal consistency reliability for the whole survey was .95. Content validity based on ratings of the correspondence between SILL items and taxonomy items (as judged independently by two raters) was .98. There was no evidence of bias in students' responses.

The factor analysis involved Promax oblique rotation and resulted in 10 factors, shown below. All the individual factors except one were rather reliable, with alphas of .62 to .87 (except for Factor 6, social strategies, which was less reliable at .31). Factors included:

- general learning strategies
- authentic language use
- searching for and communicating meaning
- independent strategies
- memory strategies
- social strategies
- affective strategies
- self-management
- visualization
- formal model-building

Cross-Validation and Additional Factor Analysis

A cross-validation was conducted at Purdue University using a sample of approximately 1,200 students (Study 2). Reliability of the SILL for this sample was .96. Another factor analysis was conducted using the same procedures as for Study 1, with the intention of determining whether the factor structure would be different for a group of students as compared to the earlier group of military personnel (Study 1). The results of Study 2

indicated that several of the factors were very similar across the two different groups, but that the ordering was slightly different. Reliabilities of the various factors in Study 2 were relatively high.

Study 2 went beyond Study 1 in determining the effects of background variables, such as motivation, sex, etc., on the choice of language learning strategies. Highly significant effects were found. For instance, more motivated language learners used significantly more strategies, and more often, than did less motivated language learners. Females used significantly more than males. Other variables also had significant effects on strategy choice.

Application of the SILL in Quantitative and Qualitative Ways

Study 3, conducted at the Foreign Service Institute, was an application of the SILL, not a new validation study. The factors found in Study 1 were simply applied in the scoring of the SILL in Study 3. The primary finding from this very rich quantitative and qualitative study was that choice of strategies was highly consistent with the individual's general learning style or orientation.

The SILL in the LSCP

The SILL is now being used in the LSCP, as was initially planned. However, different procedures for factor analysis have been employed. Specifically, a maximum likelihood factor extraction has been used, with Varimax rotation of factors. The Purdue University data base of 1,200 students was chosen as the factor analytic source, because the LSCP did not readily produce a suitable factor extraction.

An overall factor related to effort beyond the classroom, and therefore motivation, was found, with a reliability of .96. Since motivation was such an encompassing variable, it was important to control its effects somewhat in order to see more clearly the structure of other factors. To do this, the lowest fourth of the sample, that is, those who reported using language learning strategies the least often and could therefore be assumed to be the least motivated to use strategies (as shown by the overall factor) were identified. Then this lowest fourth was dropped from the sample, and the whole factor analysis was rerun using the three-quarters of the sample who could be considered more frequent strategy users and more highly motivated to use strategies. A six-factor solution was then obtained, and it was applied to the current LSCP sample (approximately 1,400 cases) to produce reliabilities as follows:

functional practice	.88
good general study habits	.84
analysis and meaning search	.82
mnemonic devices	.73
intensity and perseverance	.50
conversational input elicitation	.77

A slightly earlier, six-factor solution was used in multiple regression analyses conducted with the LSCP sample, as reported elsewhere by Lett and O'Mara. Multiple regression indicated that the use of language learning strategies significantly predicted language performance, over and above the prediction provided by other variables which were entered earlier into the regression equation. This means that language learning strategies do appear to have a strongly predictive influence on the performance of language learners, as would be expected.

Conclusions

It is possible to draw several conclusions from the information just presented. First, language learners of various types and occupations employ a wide range of language learning strategies. Second, across groups there are often similarities in factor structures of strategies, but sometimes groups differ in the ordering of these strategy factors. Finally, a number of important variables, such as motivation, learning style, and sex, help determine the choice of strategies.

Note

This paper was presented at the annual conference of the Military Testing Association in Arlington, Virginia on November 30, 1988. Development of the Strategy Inventory for Language Learning was funded by the Army Research Institute via various contracts through Battelle Laboratories, the Army Research Office, and Kinton, Inc. The SILL was developed for use in the Language Skill Change Project, an ongoing, joint effort of the Army Research Institute and the Defense Language Institute. The author has most recently served as consultant to ARI through a contract with Advanced Technology, Inc. on the subject of using artificial intelligence for foreign language learning. After January 1, 1989, the author can be reached at the following address: Director, Intensive English Communication Program, The Pennsylvania State University, University Park, PA 16802.

Identifying Precursors of Success in Foreign Language Learning

Francis E. O'Mara, Ph.D.

Advanced Technology, Inc.

In 1985, the Army Research Institute for the Behavioral and Social Sciences (ARI) and the Defense Language Institute and Foreign Language Institute (DLI) initiated the Language Skill Change Project to examine the development and retention of foreign language skills among Army personnel. This project is longitudinally examining cohorts of Army intelligence specialists to study how job relevant foreign language skills develop in these individuals as they proceed through their DLI training, through their intelligence training, and over their initial two years of service. Through analysis of the individuals' pretraining abilities, backgrounds, and characteristics; the nature of their foreign language training and experience in the Army; and their periodically assessed foreign language skills and job performance, this effort is attempting to achieve three objectives:

- To determine the presence, direction, and extent of changes in language proficiency following resident training
- To evaluate the influence of several factors (e.g., individual abilities and characteristics, post-DLI language training and experiences) on language skill change during and after training
- To determine the relationship between language proficiency and job performance.

The present report represents the initial series of analyses from this project. In its focus on the initial development of foreign language skills through DLI training, it addresses an issue related to the project's second objective: identifying those factors which predict success in initially learning the foreign language during DLI training. This success subsumes two components: successful completion of the language training, and the level of foreign language competence attained at the completion of the training.

Project Variables

Guided by the existing literature on second and foreign language learning, several types of variables can be identified as potential predictors of success in second language learning. Each of these are considered in turn in the sections which follow.

Demographic Variables

The demographic variables associated in the literature with successful foreign language learning include;

- **Level of education** -- Successful educational experience is a joint indicant of an individual's ability and motivation to successfully perform in an academic environment.
- **Age** -- There are conflicting results from studies on the effect of age upon language acquisition (Genesee, 1978, 1980; Lenneberg, 1967). Because the range in ages considered in the present effort is very small compared to that typically considered in the literature, age differences observed in the present effort may reflect maturity and ability to perform in the structured military environment characterizing the learning situation.
- **Sex** -- Sex has been shown to be an important variable in first language development (Gage & Berliner, 1975). It can be expected that these differences will extend to second language learning as well.
- **Brain hemisphere dominance** -- Neurological studies have shown that the language function is largely concentrated in the left brain hemisphere. It is a reasonable premise then that people who are left-brain dominant (i.e., right-handed) would have superior verbal skills than right-brain dominant individuals. Research has demonstrated this skill difference, suggesting that a parallel difference can be expected in second language learning.

Ability

Ability subsumes general ability, memory, and verbal ability as well as the more specific aptitude to learn a foreign language. General intellectual ability is likely to have an effect upon the acquisition of a second language as it will on any form of learning or cognitive task, and especially because of heavy verbal component typically included in the measurement of such ability.

Beyond this general ability, more specific cognitive abilities supportive of language use and development are likely precursors of successful second language skill development. The ability to make grammatically correct judgments in the primary language was found to be a predictor of second language proficiency among university students (Masny & d'Anglejan, 1985). The ability to learn and remember unfamiliar terms is also important for learning language (Mann, 1984).

Personality/Cognitive Style

A number of personality and cognitive style variables have been advanced as predictors of foreign language acquisition. These include empathy (Naiman, Frolich & Todesco, 1975), extroversion (Naiman, *et al*, 1975), field independence (Bialystok & Frolich, 1978), analytic thinking (Shipman & Shipman, 1985), tolerance for ambiguous situations (Witkin & Goodenough, 1977), and intellectuality (Kirby & Hiller, 1973).

Attitudinal/Motivational Variables

Research by Gardner and Lambert (1959) has suggested that attitude and motivation to learn a foreign language are major factors related to language achievement. Based on the decades of research conducted by he and his associates, Gardner (in Gardner, Lalonde, Moorcraft & Evers, 1985) has identified these variables as central to the foreign language learning process.

Language Learning Experience

It can be expected that past experience in learning a foreign language will be a strong indicator of success in future learning of that language. This can be due to the head start provided in the earlier training in knowledge of the language's vocabulary and syntax, as well as to the positive attitudes toward the language acquired through this training. Previous experience in learning other foreign languages should also be a predictor of learning acquisition success, both as an indicant of the desire and ability to learn languages and because of the acquisition of a broader understanding of a different syntactical structure.

Learning Strategies

A recent and growing interest in the language learning literature has been the role of learning strategies in acquiring and developing foreign language skills (Wenden & Rubin, 1987; Oxford, forthcoming). A growing body of evidence has suggested that good language learners use more strategies, better strategies, and strategies more appropriate to particular language skills (Papalia & Zampogna, 1977; Tyacke & Mendelsohn, 1986). While learning strategies are quite likely to be especially potent in naturalistic language acquisition, it is also likely that the self-initiated use of appropriate learning strategies will contribute to foreign language learning in an intensive learning environment such as DLI.

Method

All enlisted Army personnel attending DLI for Korean, Russian, German and Spanish training, beginning in February, 1986 and completing classes by July 1987 were included in this research. Those who attrited for non-academic reasons were excluded, as were those who transferred from one language course into another or were recycled into a later class during their training. A total of 1903 subjects were administered an initial battery of instruments in their initial week of language training; of these,

1383 successfully completed their training and were administered a language proficiency test by the end of July 1988. Lett presents further description of this sample elsewhere in this document.

In addition to measures gleaned from personnel records, data were collected for each subject in the course of three waves of data collection. The first two were used to collect data on measures to be used as predictors of student DLI performance. They consisted of sets of aptitude, attitude, and personality instruments, as well as a questionnaire on foreign language background administered in a group set-

MEASURE/INSTRUMENT	VARIABLE	VARIABLE GROUP	WHEN MEASURED
Records	Level of Education	Demographics	Time 1
Records	Pay Grade	Demographics	Time 1
Records	Sex	Demographics	Time 1
Background Questionnaire	Handedness	Demographics	Time 1
ASVAB GT Score	Cognitive Ability	Cognitive Ability	Time 1
DLAB	Language Aptitude	Language Aptitude	Time 1
Flanagan Expression Test	Syntax Skills	Other Abilities	Time 1
Flanagan Memory Test	Verbal Memory	Other Abilities	Time 1
Group Embedded Figures Test	Field Dependence	Other Abilities	Time 1
Critical Thinking Appraisal	Analytic Reasoning	Other Abilities	Time 1
California Personality Inventory	Empathy	Personality	Time 1
Eysenck Personality Inventory	Extraversion	Personality	Time 1
MAT-50	Tolerance of Ambiguity	Personality	Time 1
Personal Outlook Inventory	Intellectuality	Personality	Time 1
Garner Motivational Scale	Motivational Orientation	Motivation at Start	Time 1
Background Questionnaire	Prior Training in Studied FL	Prior FL Training	Time 1
Background Questionnaire	Prior Training in Another FL	Prior FL Training	Time 1
Background Questionnaire	Pre Training Proficiency in FL	Prior FL Proficiency	Time 1
Background Questionnaire	Proficiency in Another FL	Prior FL Proficiency	Time 1
Garner Attitude/Motivation Scales	Training Attitudes & Motivation	Motivation in Training	Time 2
SILL	Language Learning Strategies	Learning Strategies	Time 2
DLPT	Attained FL Reading Speaking & Listening Skills	Criterion Measures	Time 3

Table 1. Research Measures.

ting to the subjects. The first of these occurred in the first week of language training (termed "Time 1" in this paper) while the second took place six weeks after the training began ("Time 2"). The "Time 3" data collection consisted of the standard administration of the Defense Language Proficiency Test (DLPT) at the end of training. The measures taken at each of the three points in time are displayed in Table 1.

Analyses & Results

Multiple regression analyses were conducted on the collected data to determine the degree to which the predictor measures collected at Times 1 and 2 predicted success in foreign language learning at DLI. Two lines of analysis were pursued: the first examining the prediction of DLI attrition, and the second looking separately at the prediction of acquired speaking, listening, and reading skills at the end of the language training.

Since the order with which predictor measures are entered into a regression analysis can influence the analysis results, some thought was given first to the basis of this order. To optimize the cost/benefits of implementing the research results, it was decided to structure the tested prediction model in an order of increasing implementation costs. That is, the measures already being used to select and place DLI students (the ASVAB and the DLAB) were included first in the prediction equation. These were followed by measures for which data was already available, though not presently used for student selection (Demographic measures). Next were included measures for which data could be readily obtained at low cost to support selection decisions

PREDICTOR GROUP	SPANISH	GERMAN	RUSSIAN	KOREAN	ALL
ASVAB	.00	.01	.04 **	.01	.01 **
DLAB	.02 *	.01	.16 **	.08 **	.06 **
Demographics	.03	.04 *	.03 **	.04 **	.03 **
Handedness	.02	.01	.00	.01	.01 *
Prior FL Training	.07 **	.03	.00	.00	.03 **
Prior FL Proficiency	.00	.00	.61 *	.01	.01 *
Motivation at Start	.02	.03 *	.01	.00	.00
Personality	.02	.01	.01	.02	.00
Other Cognitive Abilities	.04 *	.01	.03 **	.04 **	.02 **
<i>R</i> ²	.21 **	.15 **	.28 **	.21 **	.16 **
<i>Adjusted R</i> ²	.13	.08	.25	.15	.15
<i>N</i>	238	281	516	228	1383

Table 2. ΔR^2 & Summary Statistics for Analyses on DLI Attrition.

PREDICTOR GROUP	SPANISH	GERMAN	RUSSIAN	KOREAN	ALL
ASVAB	.11 **	.14 **	.04 **	.00	.03 **
DLAB	.02 *	.02	.08 **	.07 **	.01 **
Demographics	.01	.00	.00	.02	.00
Handedness	.00	.02	.01 *	.02	.01 *
Prior FL Training	.02	.03	.02	.01	.02 **
Prior FL Proficiency	.01	.01	.00	.02	.01 *
Motiv. at Start	.03	.01	.01	.01	.00
Motiv. During Training	.05 *	.05 **	.07 **	.03	.05 **
Learning Strategies	.09 **	.07 **	.03 *	.00 *	.04 **
Personality	.01	.04 *	.02	.07 **	.03 **
Other Cog. Abilities	.03	.04 **	.02 *	.01	.02 **
<i>R</i> ²	.39 **	.44 **	.32 **	.33 **	.26 **
<i>Adjusted R</i> ²	.26	.33	.25	.21	.22
<i>N</i>	154	188	184	200	800

Table 3. ΔR^2 & Summary Statistics for Analyses on Attained Proficiency in Foreign Language Reading.

(Handedness, Prior Foreign Language Training, Prior Foreign Language Proficiency). Finally, measures which, if they were included in DLI student selection, would require administering and scoring additional instruments were entered into the regression analyses. In doing this, the motivational measures (Time 1 Motivation and Time 2 Motivation during Training) were entered first, followed by the Learning Strategies measures, the scores on the various Personality measures, and then the scores achieved on the Ability measures other than the ASVAB and DLAB. The measures falling into each of these blocks of measures are indicated in Table 1.

Table 2 presents the results of the analysis conducted on DLI attrition, while those for attained proficiency in reading, speaking and listening in the foreign language are displayed in Tables 3, 4, and 5, respectively. In each of these tables, the statistic is the increment of predictive variance provided by each block of variables and the total dependent measure variance accounted for by the prediction model.

Conclusions

The results presented in Tables 2-5 support the following conclusions:

Success in language learning at DLI can be predicted. -- Despite the rigorous standards by which DLI students are selected, there remain systematic differences between those who succeed at DLI from those who do not. With only a single exception, the twenty prediction equations tested in this effort significantly predicted DLI attrition and achieved foreign language proficiency. On the average, the significant prediction equations accounted for 28% of the variance in the criterion measures.

Predictability varies across criteria of student success -- In general, student attrition is less well predicted by the variables considered than are the measures of student foreign language skill. In part, this may be due to the fact that attrition, unlike the proficiency measures, is a dichotomous variable and hence is by its

PREDICTOR GROUP	SPANISH	GERMAN	RUSSIAN	KOREAN	ALL
ASVAB	.09	.01	.00	.02	.00
DLAB	.00	.00	.01 *	.01	.01
Demographics	.01	.04	.01	.01	.00
Handedness	.01	.02	.00	.00	.00
Prior FL Training	.03	.10 *	.02	.02	.06 **
Prior FL Proficiency	.08 **	.01	.01	.02	.02 **
Motiv. at Start	.02	.01	.00	.00	.00
Motiv. During Training	.13 **	.03	.04 **	.01 *	.04 **
Learning Strategies	.03	.04	.03	.03	.03 **
Personality	.01	.01	.01	.01	.00
Other Cog. Abilities	.02	.04 *	.03 **	.01	.00
<i>R</i> ²	.33 *	.32 *	.17 **	.18	.17 **
<i>Adjusted R</i> ²	.18	.19	.08	.04	.14
<i>N</i>	164	188	329	200	681

Table 4. ΔR^2 & Summary Statistics for Analyses on Attained Proficiency in Foreign Language Speaking.

PREDICTOR GROUP	SPANISH	GERMAN	RUSSIAN	KOREAN	ALL
ASVAB	.11 **	.08 **	.05 **	.00	.02 **
DLAB	.00	.01	.06 **	.02 *	.06
Demographics	.05 *	.00	.01	.02	.00
Handedness	.02	.02	.01	.00	.00
Prior FL Training	.05 *	.05 *	.01	.02	.04 **
Prior FL Proficiency	.06 **	.01	.01	.01 *	.01 *
Motiv. at Start	.05 **	.04 *	.02 *	.01	.01 **
Motiv. During Training	.00	.06 **	.01 *	.05 *	.04 **
Learning Strategies	.07 *	.06 *	.08 **	.04	.05 *
Personality	.01	.01	.01	.04	.00
Other Cognitive Abilities	.02	.02	.04 **	.04 *	.01 **
<i>R</i> ²	.45 **	.46 **	.30 **	.21 **	.37 **
<i>Adjusted R</i> ²	.32	.35	.21	.14	.31
<i>N</i>	164	188	329	200	681

Table 5. ΔR^2 & Summary Statistics for Analyses on Attained Proficiency in Foreign Language Listening.

nature more difficult to statistically relate to the predictor variables. Beyond this, the results may reflect the fact that the precursors of DLI attrition may fall outside even the comprehensive set of predictors used in this research.

The three foreign language skills tested at the end of training also differed in their predictability, with the prediction of attained speaking skill falling far short of that for either reading or listening comprehension. Since subsequent analysis confirmed that this difference could not be attributed to differences in the way in which the three skills are scored on the DLPT, this difference is attributable to either the inherent difficulty in reliably measuring speaking proficiency and/or the greater complexity involved in learning to generate language vs. interpreting that spoken or written by others.

Cognitive ability consistently predicts success in foreign language learning -- Where achieved foreign language proficiency was most strongly predicted (Reading and Listening), the three types of ability measures (ASVAB, DLAB, Other Abilities) figured prominently in this prediction. Among these predictors, the DLAB appeared to be more valuable in predicting success in the more difficult languages, while the ASVAB GT was more fruitful when applied with those which are less difficult. This difference may reflect the greater power of the DLAB to discriminate differences in the very high ability ranges found among students of Russian and Korean.

Non-cognitive measures offer significant potential in predicting success in acquiring foreign language skills -- Student attitudes, motivation, and applied learning strategies, were prominent in the prediction of listening and reading skills, while motivation provided relatively considerable increments of prediction to the less predictable speaking skills. This result points to the potential value of improving students' DLI performance through methods beyond more stringent selection. Further analyses in this project will be directed to identifying the particular motivational dynamics and learning strategies underlying these results. It will be possible to apply these findings to identifying and rectifying problems students consistently encounter in their language training which impede the ultimate success in that training.

NEW TECHNOLOGIES FOR FOREIGN LANGUAGE TRAINING AND SUSTAINMENT

Melissa Holland, Stan Kostyla, Merryanna Swartz, Joe Psotka
U.S. Army Research Institute

The Problem

The data on loss of language skills reported in this symposium call for ways to sustain those skills after students leave the Defense Language Institute (DLI). This paper examines the potential of new technologies for training foreign language skill and preventing its loss. To focus this examination, we isolate a critical area of foreign language knowledge - the specialized vocabulary needed for military jobs such as intelligence.

Acquiring and maintaining this specialized vocabulary is a problem for military linguists. Job-specific linguistic knowledge is not taught at DLI but is presented as part of learning center activities during subsequent job training. In the case of interrogator training, for example (MOS 97E), students must learn hundreds of new technical and military terms needed to conduct questioning in a foreign language. Typically, these words are presented in list form and students are required to memorize the lists. But the method is often ineffective. Tests given at the end of training, in which students translate interrogation scripts to and from their target language, show consistently low scores, which instructors attribute largely to students' failure to grasp the necessary military vocabulary. Moreover, indications from the field are that students rapidly forget the vocabulary they have acquired once they leave school.

The need to train and sustain this job-critical terminology is a major impetus of our current work. Sponsored by the U.S. Army Intelligence Center and School (USAICS), we are exploring how to better teach job-specific language skills by exploiting new technologies from the cognitive and computer sciences. Our work should lead to systems that can not only train in the schoolhouse but also sustain in the field, where instructors are in short supply and where realistic experiences in the target language may be sporadic.

Addressing the Problem: Conceptual Analysis

Our first step in addressing how to teach and sustain vocabulary was to ask how words are organized in memory. The organization of the mental lexicon is the topic of a body of psychological research in English - e.g., Collins and Quillian (1972), Miller et al. (1988), Forster (1976). This research reveals that memory for words is highly structured along principled logical and semantic dimensions, as manifested in the systematic links between words in memory. A prominent type of link is categorical inclusion - "canary-bird," "bird-animal," - in which the categories form a hierarchy of "is a" relations (X is a [kind of] Y). Also prominent are part-whole links - "bird-wing," "man-arm." The mental lexicon can be thought of as a network of words with multiple pre-existing interconnections that follow consistent semantic dimensions.

These pre-existing interconnections are important for storing and retrieving words, as revealed in word recall studies and word association tests (Tulving and Donaldson, 1972; Deese, 1965). Words being recalled or associatively produced tend to come together in clusters that reflect semantic dimensions. For example, words referring to animals might cluster separately from words referring to minerals, regardless of the order of words in the original recall list. This clustering arises because words with prior memory links serve as retrieval cues for each other.

Semantic interconnections are equally critical for natural language use. Words strongly linked in memory tend to prime or activate each other (Forster, 1976). For example, encountering the word "doctor" in a text primes memory for related words like "nurse" and "hospital" and makes subsequent recognition of those words easier and faster. This priming is critical in reading and producing language, because semantically related words tend to co-occur in text and discourse. Thus, being primed for semantic associates strongly constrains the reader-listener's expectations about "what comes next" in a sentence or utterance. Being able to predict or limit what comes next has been shown to vastly increase intelligibility of speech and comprehensibility of text.

Given the importance of a structured lexicon for native speakers' memory for and use of words, acquiring this structure can be seen as a rational and necessary goal for foreign language learners. Not only must learners grasp the meaning of a new word in the strict sense of definition, external referent, and rules of use, but they must also integrate the word with other new words and with the existing lexicon. To be effective, this integration should follow the conventional logico-semantic organization found in native speakers. Arguably, then, vocabulary training should introduce words in the same organized form that they exist in memory.

Addressing the Problem: Hypertext Tools

Having identified the need to teach lexical structure, our next step was to find suitable software tools. Given the multi-dimensional nature of lexical connections, and the multi-modal nature of word representations in general (written, pictorial, auditory), we saw a hypertext system, like MacIntosh's HyperCard or Xerox's NoteCards, as a natural instructional medium. Hypertext software allows a designer to create collections of informational nodes of different types and at different levels and link or cross-reference them, forming a web of facts or concepts through which an enormous variety of routes is possible. The links convey the relationships that exist among the concepts. The linked nodes can then provide an environment for exploratory learning, in which students trace routes open-endedly, jump between levels, and move in different directions. Alternatively, the linked nodes can be used as the underpinning for more directed instruction, in which a student's path is guided and branching hinges on analysis of the student's responses.

In addition to cross-referencing nodes of information, hypertext is uniquely suited to linking nodes in several media, such as text, graphics, and sound. For example, the written form of a word can be linked with an animated graphic or a pronunciation, which the student accesses by touching (or buttoning) the spot on the screen where the word is displayed. A screen can have several such "hotspots" that open up further modes of representation when activated. Multi-media representations such as this constitute a basic

building block of vocabulary teaching. In addition, the capability to tie the representations together in patterns allows a hypertext database to mirror rather directly the structure of lexical knowledge, with its semantic domains and its web of connections within and between domains.

The most ambitious attempt to build a computerized lexical network is the continuing work of George Miller and associates (Miller, Fellbaum, Kegl, and Miller, 1988), who are developing an on-line reference system, called WordNet, designed to capture the organization of word memory in English. WordNet links words according to the conceptual relations of class inclusion("is-a"), part-whole, antonym, and synonymy. These four relations figure saliently in psycholinguistic studies of word memory, recognition, and usage. Equally salient in linguistic analyses, they occur pervasively in the surface structure of written and oral discourse. For example, subordinate-superordinate shifts are common in question-answer routines in everyday conversation as well as in interrogations. Among the applications envisioned for WordNet is an instructional aid in hypertext that displays conceptual links between words so that a learner can freely peruse them.

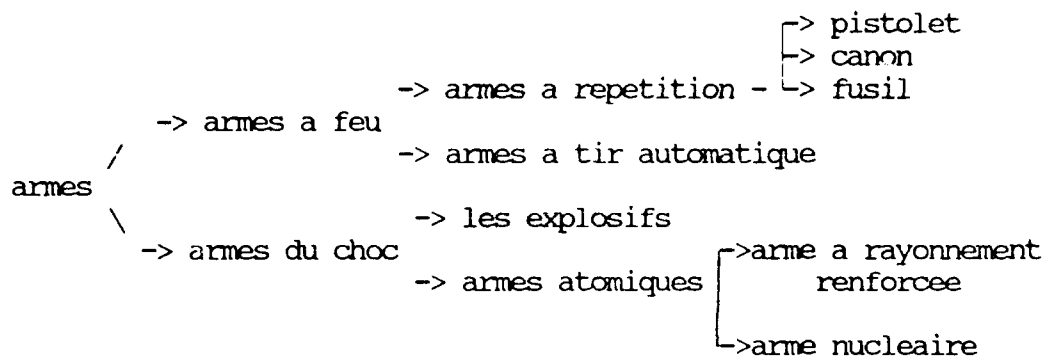
If the WordNet method were used to construct networks in other languages, learners of these languages could browse the networks, or do exercises with them, and thereby - we hypothesize - acquire, practice, and strengthen important lexical linkages. Prior evidence from presenting semantically organized mappings of words suggests that this method may be very effective for word learning in foreign languages (Carter, 1987; Cornu, 1979; Meara and Ingle, 1986). This method approaches an immersion technique, relying on word configurations in the foreign language while bypassing native-language mediators. The use of native-language mediators, as when learners study lists of foreign and English word pairs and encode the pairings, appears to promote translation strategies and curb the development of second-language fluency (Fuentes, 1986). Fluency builds up most rapidly through immersion in the second language, which fosters direct connection of utterance with meaning and inhibits the mental step of translating to and from the native language (Hamburger, 1988).

Illustrating the Power of Hypertext: Three Prototype Systems

To illustrate the potential of hypertext to create immersion style environments that display inter-word ties, we describe complementary aspects of three systems being developed at the Army Research Institute. Each system builds on Miller's WordNet idea and on the extensive psycholinguistic research that supports it.

Word trees. The first system is based on a relatively simple idea - to display the network of "is-a" relationships that characterize the military lexicon used in interrogation. We took the list of several hundred specialized terms given to 97Es and grouped them into semantic domains, like terrain features, officer ranks, and weapons and firearms. Most of the domains feature nouns, but some contain verbs, like those referring to information gathering or terrorist activities. We attempted to create "is-a" hierarchies with the words in each domain. Many domains lent themselves gracefully to this kind of analysis, and these became the data for a WordNet-like learning environment. We are translating the "is-a" hierarchies into selected foreign languages and implementing them as hypertext word trees using the Xerox NoteCards system.

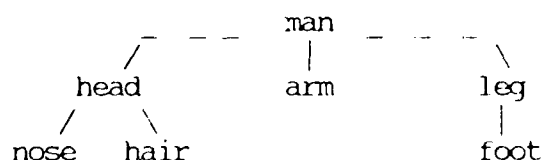
NoteCards displays the subordinate and superordinate links surrounding any word the user chooses, as seen below in the partial configuration of French words for weapons ("armes"):



In addition to displaying these links, NoteCards permits the user to go back and forth between the tree and other "cards" of information that pop up like windows inset on the screen. For each word in a hierarchy we created a card that gives a set of synonyms, some grammatical points, a sample sentence illustrating word usage, and an English translation. Users who want to see this extra information can choose it by buttoning a word in the hierarchy. Several cards could appear on the screen at once depending on the type of information the user requests.

At this point we have created a browser rather than directed exercises. We will try out the system with students at USAICS to see how well a flexible, learner-controlled environment works for advanced language learners. The next two systems couple learning by exploration with more directed, interactive instruction.

HyperLexicon. A small prototype system known as HyperLexicon displays networks of common words in English. It was designed to demonstrate the capabilities of MacIntosh's HyperCard to couple verbal with graphical representations of meaning. The initial meaning of a word is given by its place in either an "is-a" or apart-whole hierarchy, based on the WordNet principle. The words below, for example, are shown in part-whole relation to "man":



To explore analogous concepts, the user can view simultaneously three other hierarchies on the screen - showing the parts of animals corresponding to the parts of a man. The user can even work across hierarchies by pairing parallel words, like "foot" and "hoof," and get feedback on whether the parallel holds. By buttoning on a word in the hierarchy, the user can see a still or moving picture depicting the referent. For words in the set of analogous hierarchies, the picture appears in a set of four, at whatever level of the hierarchy the user indicates - thus, for "foot" there would be illustrations of a foot, hoof, paw, and (bird's) foot in each of the screen's quadrants. These pictures can

be fairly easily constructed by using HyperCard's library of graphical icons or by digitizing existing hard copy images.

HyperLexicon also has a segment integrating verb and noun hierarchies - such as verbs of eating (e.g., "ingest," "lap," "dine") and animate nouns (e.g., "animals," "cats," "people"). The user can pair verbs and nouns across hierarchies and get feedback on the appropriateness of the selection. The feedback is aimed at teaching and exercising knowledge of selectional restrictions, that is, the semantic and syntactic constraints on the kinds of subjects and objects verbs take (thus, "eat" takes an animate subject and "dine" requires a specifically human subject, except in metaphorical uses). Knowledge of selectional rules is a core component of a native speaker's competence.

As a final instructional device, motion verbs in HyperLexicon - like "walk," "run," and "swim" - are arranged not hierarchically but in a geometric network in which spatial distance indicates semantic distance. Thus "walk" and "run" are set close together and "swim" is set farther away - another method of conveying meaning relationships iconically. Using the HyperLexicon structure, we can easily translate the word tokens into a second language and try out the devices of analogous trees, pictures, and semantic geometries with language learners.

LexNet. A more advanced system called LexNet in Situ integrates semantic hierarchies like those above with realistic sentence and paragraph contexts. Applied to the domain of French words for military intelligence, this system is also implemented in HyperCard. Potentially unfamiliar words are first encountered in paragraphs that resemble a real intelligence report. When the user buttons a word, a pop-up menu appears with options of hearing the word, seeing a new sentence context, getting an English translation, or seeing a graphical representation that shows where the word fits in a network of related words. These networks organize words not only according to "is-a" hierarchies but also according to synonymy and antonym - two relations central to the cohesive structure of paragraphs and critical to a native speaker's fluency. To build automaticity in using these relations, the learner can rapidly alternate between synonym and antonym trees for a given word.

After traversing the various network structures given by LexNet, users can practice forming their own networks. The user first chooses to deal with a particular relation, like synonymy, and is then shown a group of words strewn on the screen, some of which are synonyms. The user employs simple HyperCard facilities to move words across the screen and draw links between them. Although the system is not sophisticated enough at this point to give feedback on whether synonyms are connected correctly, this capability is well within the limits of Hypercard. At this point, the system prints out all the user's responses and a teacher or fellow student can check them. In fact, this form of evaluation is required for another LexNet exercise: writing your own sentences using words from the trees and paragraphs.

A final type of exercise is the common cloze procedure, in which the user sees a paragraph with occasional blanks and fills in the blanks with words from a pool listed at the bottom of the screen. The program delivers feedback on the correctness or appropriateness of responses.

The aim of this array of exercises is to coordinate lexical knowledge with its use in text and talk. Learners are expected both to acquire the relatively abstract associations we have called lexical structure and to gain facility in retrieving and using this structure for the concrete tasks of writing, speaking, and understanding. The user is intended to build automaticity of word linkages through exposure to and practice with word networks and flexibility of linkages through observing and using the words in varying contexts. Choosing appropriate comparison interventions, we will assess the effectiveness, usability, and acceptability of the finished system with military linguists.

Our modus operandi in developing these systems is to find out what aspects of language are vulnerable to loss and then to tap what linguistics, discourse analysis, and cognitive science have to say about these aspects: what do native speakers need to know to be competent in them? We then exploit relevant technology to design ways to teach or sustain this knowledge given limited chances for instructor interaction. Future efforts will apply artificial intelligence tools such as natural language processing to analyze, diagnose errors in, and respond to student discourse.

RESEARCH ON TRAINING WITH SIMULATED NETWORKING TECHNOLOGY (SIMNET)

Dr. Jack H. Hiller, Chair
U.S. Army Research Institute
Alexandria, Virginia

This panel presents the goals and research approaches for designing a comprehensive training and evaluation program for SIMULATION NETWORKING (SIMNET) technology. This technology promises to expand the opportunities units may have for collective training (principally, command and control) at platoon, company and battalion echelons (and perhaps higher). The panel starts with a discussion of the research opportunity inherent in SIMNET for examining issues related to (1) developing a performance measurement system for SIMNET, (2) formulating training strategies for SIMNET that maximize training benefits to units, and (3) building a global training strategy that integrates SIMNET into the total Army training system. The remaining papers present major results achieved on primary program objectives. The panel concludes with a discussion of the research program from a behavioral research and combined arms training perspectives.

THE SIMNET OPPORTUNITY FOR RESEARCH ON TRAINING AND PERFORMANCE MEASUREMENT

JACK H. HILLER

U.S. ARMY RESEARCH INSTITUTE

1. Introduction. SIMNET technology is highly promising, but there are certain topics and issues that merit discussion, particularly concerning how SIMNET training may be related to the current best simulation for training heavy forces, the National Training Center, NTC:

- a. Army Goals of Home Station, NTC, and SIMNET Training;
- b. Techniques for Measuring Unit Performance;
- c. Invalid Use of SIMNET and NTC Performance Data for OERs;
- d. Use of NTC to Validate Experimental SIMNET Training;
- e. Value of Procedural Task Performance Measurement;
- f. Impact of Differences between NTC and Combat;
- g. Grand Training Strategy

Each of these topics is briefly discussed below to set the stage for this session.

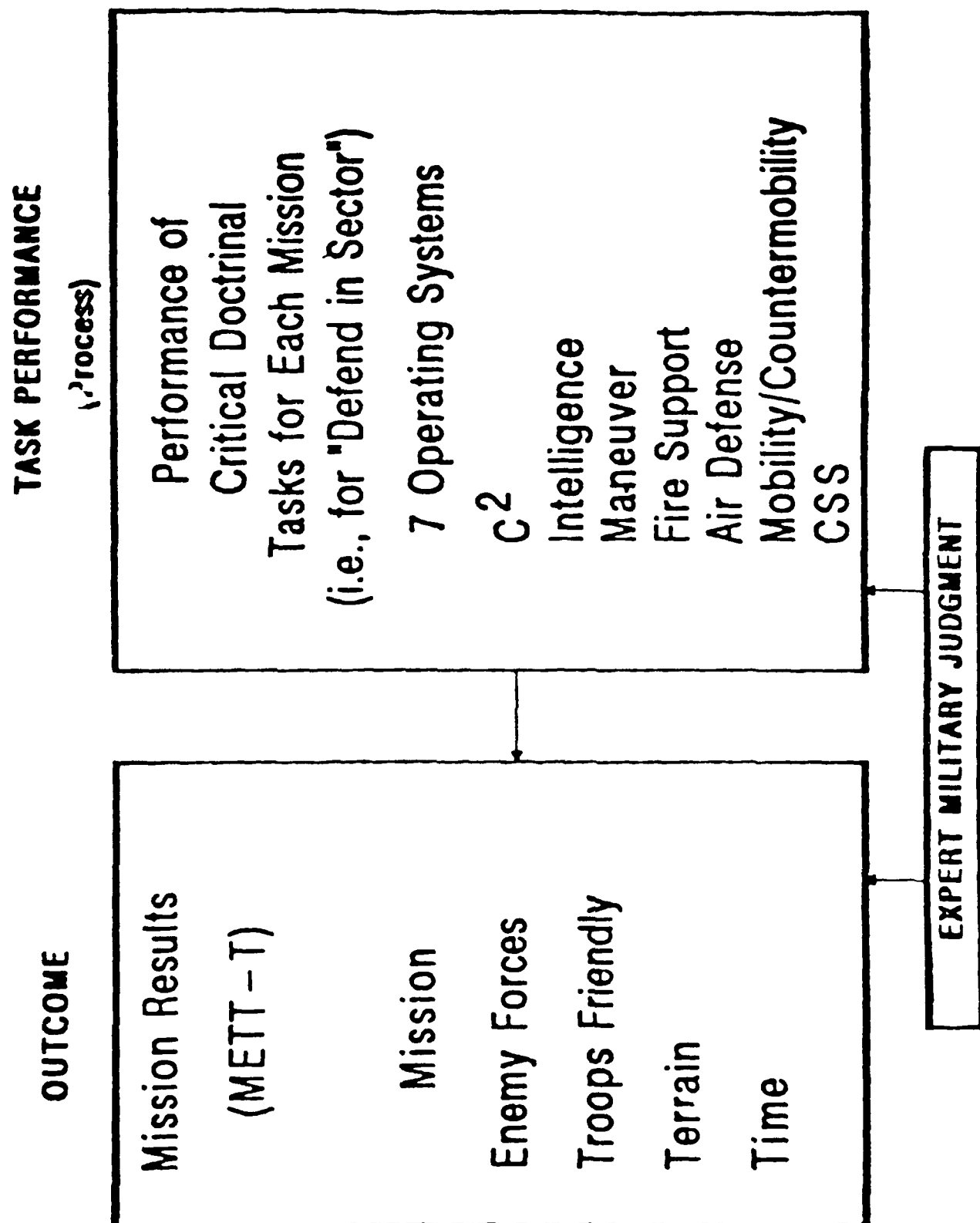
2. Goal of Army Unit Training. FM25-100 (Draft) states that the chain of command is to focus on its Mission Essential Task List (METL) as a strategy for managing training in the face of all of the possible time-demanding tasks confronting a garrison Army. NTC, SIMNET, Home-Station field training exercises (FTXs), etc., simply offer different ways of conducting training to achieve/maintain combat proficiency by focusing on each unit's METL. Successful performance at the NTC or in training conducted on SIMNET would be a valid indicator of a unit's combat preparation only to the extent that:

- * the missions/tasks included in SIMNET and NTC correspond to the unit's METL;
- * the mission scenarios faithfully reflect the conditions of combat.

3. Techniques for Measuring Unit Performance. ARI has proposed a three dimensional model, at Figure 1, for measuring performance (Hiller, 1987). Briefly stated, the three measurement dimensions are:

- * Battle outcomes, as measured by enemy and friendly attrition, and terrain control using the Multiple Integrated Laser System (MILES) for objective scoring;
- * Task performance scored for conformity with Army Training and Evaluation Program and Army Mission Training Plan (ARTEP/AMTP) procedural task standards;

FIGURE 1. THREE DIMENSIONAL MODEL FOR UNIT PERFORMANCE MEASUREMENT



- * Expert judgment keyed to major performance dimensions, such as the operating systems, using rating scales, and with explanations required for ratings departing from average scores.

Measures produced with this model should have the potential for providing trainers with enhanced performance feedback and researchers with information useful for determining possible Lessons Learned in training and doctrine.

4. Application of Unit Performance and Performance Effectiveness Scores to OERs. Application of any of the three measurement dimensions to score any unit's performance on a mission conducted on SIMNET or at the NTC will not necessarily produce measurement of the unit's performance effectiveness. The reason derives from variability in conditions under which the mission was performed: for example, fighting the NTC OPFOR vs another unit from the same division. Under no circumstances should the research community argue for standardized performance conditions where units are training to mastery levels, because of the need to learn how to deal with the surprise and the "fog of war." Because of natural and controlled variation in conditions of performance difficulty, scoring a unit's performance effectiveness is complicated and inherently unreliable. Mechanical use of performance scores for rating unit and leader effectiveness is intrinsically invalid. Mechanical use of performance effectiveness scores is suitable for research and training feedback but, because of uncertain reliability, would not meet legal or ethical requirements for fairness if applied directly to Officer Efficiency Ratings (OERs). However, the chain of command necessarily considers unit training performance when evaluating leaders but could not and should not attempt to use scores out of context.

5. Use of NTC Performance to Validate Experimental SIMNET Training and Feedback Designs. The issue is the need to check the value of experimental SIMNET training program designs against an external criterion. The strategy selected is to see if units having had SIMNET in their train-up do better at the NTC than other units (for research purposes only). A methodology for measuring and comparing unit performance is found in the three dimensional ARI model described above. The application of the outcome measurement scheme is straightforward. However, implementation of the methodology for measuring task performance requires further explanation.

6. Value of Procedural Task Performance Measurement. Research conducted by ARI, with BDM support, over the past three years under Combined Arms Training Board (CATA) and NTC sponsorship has created an expanded, refined ARTEP design which will serve as the model for future TRADOC school and center ARTEP development. Many of the procedural tasks in the refined ARTEP are equally and highly relevant to NTC and SIMNET training. Go and NO GO

scores have a direct, meaningful interpretation for training, e.g., "Platoon elements always remain within supporting anti-armor direct fire range of one another." SIMNET technology offers a tremendous opportunity to observe, measure, record, and provide training feedback with reliability and precision unattainable out in the field. For example, platoon distribution of fire, command and control communications, and the opening time for firing (trigger-pulling time) can be precisely and easily found in SIMNET, but they are very hard to determine during simulated combat on real terrain.

7. Impact of the Limitations in Realism of Combat Simulated at the NTC. NTC has been selected as the criterion or "ground truth" for evaluating experimental SIMNET training simply because it is without challenge the most realistic combat simulation for the combined arms routinely available. However, the NTC has limitations in simulating features of combat that SIMNET can represent. For example, SIMNET may play artillery, close air support and air defense more realistically than the NTC. It is therefore possible that command, control, and tactical decision making skills learned in SIMNET would not be fully exercised and measured at the NTC. In this regard, important skills acquired while training with SIMNET might not be adequately reflected when measuring performance at NTC to test the effectiveness of SIMNET training. It is also conceivable that units could spend too much time on SIMNET and thereby fail to acquire those skills that SIMNET does not exercise, e.g., digging-in, camouflaging, and emplacing obstacles. However, any such deficiencies would be reflected by NTC performance measures.

8. Grand Training Strategy. It is possible that units who do not train enough on real equipment out in the field may not be able to capitalize on the command and control skills gained in SIMNET. We can expect some balance of weapons systems live fire and simulation training, small unit field training, and company team and battalion task force command/control training with SIMNET to be optimal; but we currently lack any empirical data to validate strategies for allocating training time. Research described in this session should eventually progress from design of theoretical models to empirical tryouts and validation.

REFERENCE

Hiller, J. H. (1987). Deriving Useful Lessons from Combat Simulations. Defense Management Journal. 2nd and 3rd Quarter, 28-33.

THE SIMNET RESEARCH PROGRAM: Overview and Performance Measurement System Progress

Richard Vestewig, Ph.D.
Perceptronics, Inc.
21135 Erwin Street
Woodland Hills, CA 91367

Introduction

DARPA's Large Scale Simulation Networking (SIMNET) program is a breakthrough concept in training. Present plans are for the production of approximately 250 simulators to be installed at sites in the US and in Europe. The capability to network these simulators makes possible training in unit coordination and command and control that hitherto has been achieved only during exercises at the National Training Center.

SIMNET also offers unprecedented opportunities for research into weapon system design, training effectiveness, performance measurement, and training system management concepts. ARI's program, "Unit Performance Measurement and Training Program Design for Networked Simulators", addresses the last two areas. Most simulators at best allow research on transfer from the simulator to the actual equipment. Although this is important, a battle is won or lost on how each weapon system acts in concert with all others. Being able to objectively evaluate coordinated action means that tactics, mission, and utilization can all be examined in their most relevant context. Dealing in this larger context makes possible new applications of performance measurement and training program integration.

Although SIMNET is objectively a set of networked part task trainers, conceptually it has much greater scope. Each simulator is a "window" onto the battlefield. The CIG world is the arena for all action and interaction. Much more than simulating the physical weapon system, SIMNET simulates the battlefield; the simulator itself is almost incidental. In addition, the entire SIMNET complex makes up the rest of the simulation, since planning, preparation, and After Action Reviews also take place there before and after the simulators engage each other in battle. In this larger context SIMNET is much more akin to the National Training Center than it is to a trainer, since the NTC provides a common ground for each participant to interact with all others.

The value of SIMNET for the research world lies in the control and data collection capabilities. Vehicles can be placed anywhere on the terrain data base, and their movements monitored in real

time. Information passed over the network to control the simulation provides location and fire events, so that each shot can be exactly traced from its inception to its result. Moreover, important information on vehicle status, such as initial fuel and ammunition stores, provide unit commanders and exercise evaluators with insight on these how resources are managed in a coordinated battle.

Many of the data passed over the network have analogues in the data collection capability at the National Training Center, especially data on vehicle location and firing events. The limitations of MILES lead to ambiguities in determining the effect of each round, so that accurate determination of fire is much less possible at the NTC than in SIMNET. However, communications are routinely collected at NTC, whereas they are not at the SIMNET facility (although inclusion of this is being implemented). Design of a performance measurement system for SIMNET should have maximal commonality with that available at NTC so that performance can be compared.

Although NTC and SIMNET have impressive data collection and recording capability, each is still only a simulation of battle. Each has artificialities that make suspect any claim that fidelity is high enough to reflect all contingencies of the actual battlefield. Therefore, conclusions drawn from research in each of these arenas must be tempered with consideration of how the settings differ from actual battle. NTC, for example, does not simulate well indirect fire or close air support. SIMNET does not as yet simulate infantry, allow terrain modification, or require that commanders deal with casualties. If these conditions change how a commander performs, then transfer to actual battlefield conditions, or even between exercises on SIMNET and NTC, will be lessened. Any research program involving these simulations must clearly understand their limitations so that conclusions can be appropriately drawn.

Goals of the Research Program and Performance Measurement System

ARI's "Unit Performance Measurement and Training Program Design for Networked Simulators" has ambitious goals. Performance measurement entails the development of a data collection, analysis, and interpretation capability, using to the extent possible off the shelf software and hardware compatible with US Army standard, and building upon work underway at the National Training Center. Training program design addresses the use of SIMNET as a training environment and how SIMNET fits into an overall training management system. In addition, the program examines a possible global training strategy and management system for the upcoming Close Combat Tactical Training (CCITT) system. CCITT is to be developed using the SIMNET concept, and is being procured by PM TRADE.

The success each of these objectives requires close coordination with various activities in the user

community as well as the research community. Representatives from PM TRADE, CATA, TRADOC, and the Armor School are attending periodic program reviews to ensure relevance and useability of the products.

The performance measurement system for SIMNET will key on relevant ARTEPs and AMTPs to provide a framework for evaluating performance. The specific focus will be the 2A Experimental AMTP being developed for ARI. 2A addresses training objectives to support training and evaluation of units at the National Training Center, and uses standard mission scenarios, such as Hasty Attack and Defend, as organization templates. The performance measurement system will be based on a software architecture which collects and analyzes data carried on the SIMNET network, and on inputting observer evaluation data for those tasks which cannot be collected automatically. One of the critical tasks, underway now, is determining the degree to which data can be collected automatically from the SIMNET network, and how much observer/controller input is required.

The primary goal of the performance measurement system is to help the unit commander by allowing him to review the results of his unit's exercise in SIMNET on hardware available at his home station, and plan additional training accordingly. Ideally, if SIMNET is used before a rotation at NTC, the unit can potentially practice and perfect skills so that the exercises at the NTC are more beneficial. However, the system will also have additional statistical analysis and graphics capability to allow the researcher to perform more sophisticated analysis. Finally, the system will have data formats compatible with NTC archives so that analysis can be performed on each data set using the same tools.

How best to use SIMNET as a training environment is of paramount importance. SIMNET is presently used both in free play, manned force on force exercises to simulate the unpredictability of actual opponents, and in exercises developed by the unit commander to practice particular skills. DARPA is also developing a semi automated OPFOR to fill out the battlefield and provide opponents with particular characteristics. This program is developing a prototype training strategy and management system for SIMNET so that each exercise meets specific training goals. Important questions being addressed are the identification of prerequisite skills so that positive transfer to operational equipment occurs and negative transfer (especially potentially dangerous performance such as driver safety errors) does not occur; scenario design to meet training goals at particular points in a learning curve; and feedback and control guidelines for exercise management.

The final goal of the program is the integration of SIMNET, and the eventual CCTT, into the total set of training devices, exercises, and decision aids for the unit commander and training manager. SIMNET is not intended to replace any training system; indeed, as a command and control training system, it assumes that basic procedure and skills on equipment operation are provided by other

training capabilities. However, SIMNET's unique features as a battlefield simulation, and its intended broad distribution, requires knowledge of its strengths and weaknesses in training so it can be well integrated with live fire exercises, TEWTs, STXs, UCOFT, and command group battle simulations. Sequencing, time spent, and remediation for SIMNET must be addressed. In addition, as the Army develops and installs the Integrated Training Management System (ITMS), SIMNET and CCTT must be integrated into the total training management capability for which ITMS aids training decisions.

The remainder of this paper will address progress in the development of the SIMNET Performance Measurement System and its features. Other papers will address progress in development of training management, 2A insertion, exercise planning, and user coordination.

Progress on Performance Measurement and Software Package Development

The Unit Performance Measurement program has a number of software deliverables focusing on automated data collection, performance measurement, and training management for SIMNET. The goal is a modular software architecture so that all software components can be accessed from a common interface. This development philosophy reduces the likelihood of redundancy, or the necessity to modify earlier modules to account for later development.

The heart of the software is the SIMNET performance measurement system. Our goal was to be as responsive to two important user groups. The first is the unit commander, who needs to review and replay SIMNET exercises to help determine areas of good and bad performance, and suggest additional training needs. The second is the researcher, who requires more sophisticated data analysis tools and the ability to analyze data from at least two sources, namely SIMNET and the ARI NTC Archives.

Equally important as user relevance is an accessible hardware and software platform. The SIMNET data capture and analysis package is PC/AT based, compatible with the standard military personal computer, the Zenith 248. Data can be directly collected from the SIMNET local area network, and stored on hard or floppy disk. They can be replayed or analyzed using dedicated functions, or an off the shelf statistical package. Data are stored in a relational data base compatible with the ARI NTC Archive data.

The SIMNET Network Protocol Data from the SIMNET local area net consist of digital protocols conveying vehicle appearance, status, firing events, and other data which portray each vehicle to others on the network. The majority of protocols are Vehicle Appearance packets, which give

information such as exact location, relative gun, turret, and hull location; role (friend or foe), company, and bumper number; velocity, engine speed, and velocity vector. Vehicle Appearance packets are broadcast on the net at a minimum of once every five seconds and a maximum of once every 1/15 second. Consequently, at the simplest level of analysis, collecting consecutive Vehicle Appearance packets allows tracking each vehicle over the terrain data base. Since turret azimuth and elevation are also shown, the Vehicle Appearance Packet also determines successful or unsuccessful fire.

The next most common data packets denote firing events. These include the Fire packet, which indicates the firing vehicle, the target, and information about ammunition fired; and Vehicle Impact or Ground Impact packet. Again, at a simple analysis level, all firing events by a vehicle can be tracked, as well as any damage caused by another vehicle's fire.

The data retrieval system simply collects all of the data packets and stores them in hexadecimal format. To date, we have successfully collected data both at the mini-network at Perceptronics and at the SIMNET facility at Fort Knox. Although all data are collected, only the four data packets described above are presently stored, since these appear to have the most relevance for performance measurement. Approximately 6 hours of a nominal 20 vehicle exercise can be stored on a 40 MB disk. Using data formats which display each packet in "human readable" form, we have tracked the history of any designated vehicle, including its movement and any fire engagements with other vehicles on the net. Note, however, that using these data for measurement of platoon and company effectiveness requires aggregation of the performance of individual vehicles.

Hardware and Software Components Hardware components of the system consist of a PC/AT or compatible, preferably with a 40Mb hard disk; a XLAN network analyzer board (PC slot compatible); monitor and printer.

Software consists of MS DOS; MS Windows for interface and control design; XQL relational data base (compatible with NTC data storage); XLAN data collection software (supplied); off the shelf statistical analysis package; and contractor developed data display format, control interface, and mission summary data adapted from NTC AAR formats.

Commonality with Other Data Bases Clearly, the ability to collect data in and of itself does not make a useable performance measurement system. Required are indications of performance that relate to successful or unsuccessful outcomes in battle. This section addresses two efforts which relate SIMNET data to important data bases and performance templates.

The four most common SIMNET data packets, accounting for over 95% of total network traffic, provide information on vehicle appearance and fire events. Although this information may appear limited, these data form the basis of data collected during NTC exercises. The results of a comparability analysis between the NTC Archive data base and the SIMNET data packets shows that these four packets are analogous to over 80% of the variables in the NTC Archive data base. Our analysis shows that collecting an additional four packets will account for the remainder.

Note, however, that this high degree of commonality does not indicate that we have a useful performance measurement data set, but rather that each data set is limited to the constraints of data collection at both SIMNET and the NTC. An additional analysis underway is the degree to which tasks and measures of performance in the 2A Experimental AMTP can be collected and analyzed automatically by this hardware/software from the SIMNET network. To the extent they cannot, observers and the unit commander may be necessary to provide the evaluation. The data collection and analysis package has the provision for inputting observer data and evaluating it against the measurement scale developed for 2A.

To the extent that all phases of the mission are evaluated - Planning, Preperation, and Execution - observer data will be necessary, since the Planning and Preparation phases take place outside of the simulators themselves. To reduce the amount of observer data and hence the number of observers, the performance measurement system may consider the determination of critical events whose performance varies strongly with other events and exercise outcome.

Conclusions and Next Steps

At this point in the program good progress is being made in the development of the performance measurement system for SIMNET. Still to be addressed are the integration of an exercise planning and control function which will aid the unit commander in setting up SIMNET exercises relevant to his training goals; and integration with the training management function of ITMS to provide an overall context for performance evaluation and training program design. We believe that succes in these areas can lead to better trained soldiers for the combined arms battlefield of the present and future.

Global Training Strategy For The Simulation Network
Jim L. Madden
The BDM Corporation

The purpose of this paper is to describe the underlying models which provide the foundation for a major portion of the Army Research Institute's (ARI) research program in support of Unit Performance Measurement and Training Program Design for Networked Simulators.

The specific focus of the paper is on the following subtasks:

Subtask 1b: Analyze potential Simulation Network (SIMNET) concepts and application and develop recommendations for a global training strategy for SIMNET placed within the larger Army training system generally and the Integrated Training Management System (ITMS) in particular;

Subtask 2c: Analyze linkages between ITMS and a SIMNET management system and identify implications for functional requirements for system to prescribe training events, user interface requirements, and ITMS/SIMNET compatibility; and

Subtask 3c: Analyze implications of ITMS maturations for the Close Combat Tactical Trainer (CCTT) training management system and make recommendations for functional requirements for CCTT training management system, user interface requirements, and CCTT/ITMS compatibility.

Due to the fact that this research program is still in its early stages, the preliminary models and systems described in this paper have not as yet been approved by either ARI or the Army. As such, at this point in time they represent only the views of the author and his colleagues.

The development of a global training strategy for placing the Simulation Network and its follow-on Army production version, the Close Combat Tactical Trainer, within the larger Army training system first requires the definition of this larger system.

While the Army has developed numerous models for unit training systems, the latest being the training management cycle in FM 25-100, Training The Force, it does not currently possess a comprehensive model which describes its larger system. Such a model can be derived, however, by starting with the Army's initial model, developed in the early seventies by the Board For Dynamic Training, and updating it in accordance with subsequent changes which have occurred in Army training systems and policies.

The crux of such a system is depicted in Figure 1.

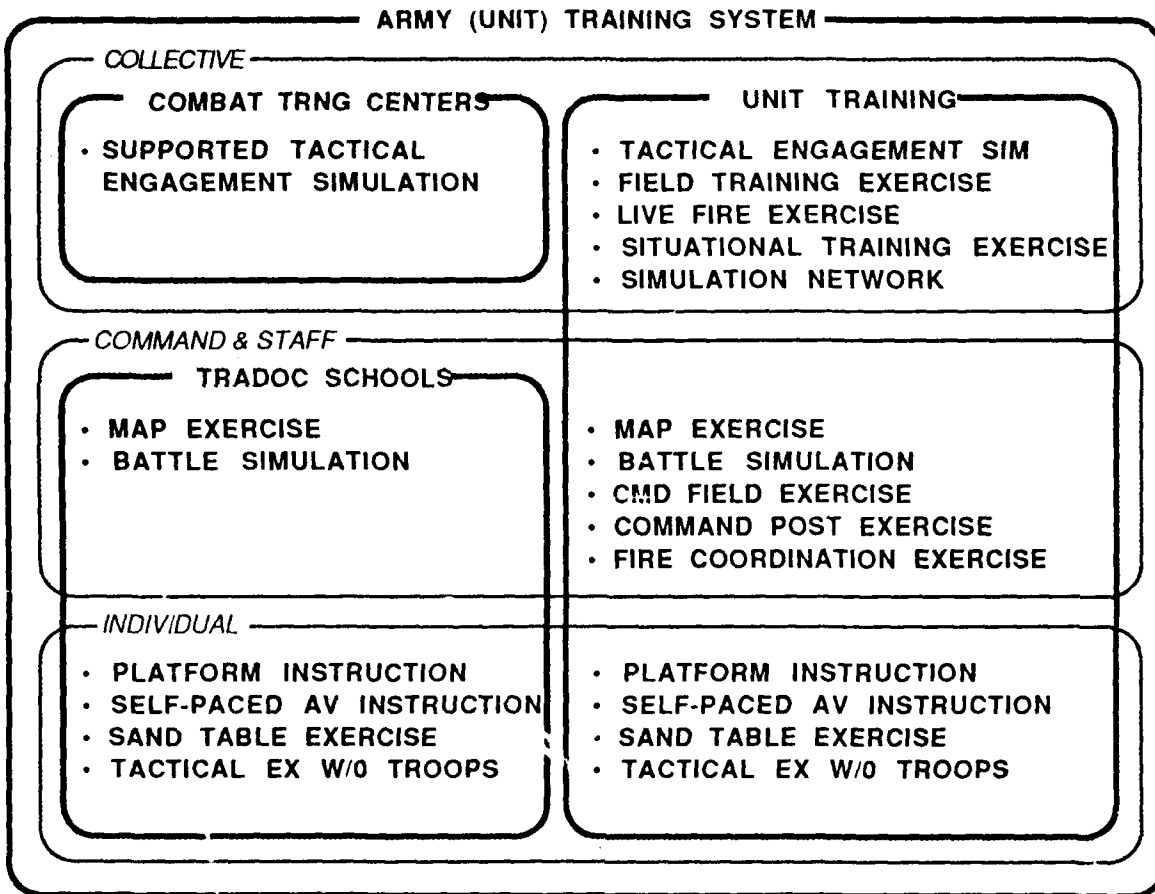


Figure 1 - Extract From Global Training Strategy

It places all of the Army's contemporary generic training systems within the context of institutional training (depicted on the left) and unit training (depicted on the right). It further breaks these training exercises down with regard to those which support collective, command and staff, and individual training. Since the types of training exercise involved are a function of the echelon being addressed, it should be noted that the above diagram is focused on the battalion level. As noted earlier, the specific placement of the various exercises depicted is preliminary and has not yet been reviewed by appropriate Army agencies.

An inspection of Figure 1 indicates that Simulation Network supported exercises have been identified as only supporting collective training within units. This has been done as the result of a preliminary analysis which indicates that the replacement of traditional forms of collective training in units with SIMNET exercises provides a potential payoff which is several orders of magnitude greater than any other possible application.

A portion of this analysis is depicted in Figure 2.

	TM MBR	TM LDR	SQD LDR	PLT LDR	CO CDR	STF OFC	BN CDR
COLLECTIVE							
SUPPORTED TACTICAL ENGAGEMENT SIMULATION	●	●	●	●	●	●	●
TACTICAL ENGAGEMENT SIM	●	●	●	●	●	●	●
FIELD TRAINING EXERCISE	●	●	●	●	●	●	●
LIVE FIRE EXERCISE	●	●	●	●	●	●	●
SITUATIONAL TRAINING EXERCISE	●	●	●	●	●	●	●
COMMAND & STAFF							
MAP EXERCISE					●	●	●
BATTLE SIMULATION					●	●	●
CMD FIELD EXERCISE				●	●	●	●
COMMAND POST EXERCISE					●	●	●
FIRE COORDINATION EXERCISE				●	●	●	●
INDIVIDUAL							
PLATFORM INSTRUCTION	●	●	●	●	○	○	○
SELF-PACED AV INSTRUCTION	●	●	●	○	○	○	○
SAND TABLE EXERCISE	●	●	●	●	●	●	●
TACTICAL EXERCISE W/O TROOPS	●	●	●	●	●	●	●
<input checked="" type="checkbox"/> High Payoff <input type="checkbox"/> Moderate Payoff <input type="checkbox"/> Low/No Payoff							

Figure 2 - Extract From SIMNET Application Analysis

The purpose of this analysis was to identify the highest payoff applications for SIMNET exercises in order to provide a focus for the research effort. While the limitations on the length of this paper do not permit a discussion of how the above findings were arrived at, a few comments may be of interest to the reader. It will be noted that SIMNET has not been indicated as an acceptable substitute for a Battle Simulation or Command Post Exercise even though considerable work is ongoing to enable SIMNET to fulfill this role. This has been done for two reasons. First, as noted in Figure 2, command and staff exercises have been intentionally developed within the Army's multi-echelon training philosophy to support the training of those personnel without the participation of subordinate units. The incorporation of SIMNET Combat Vehicle Simulators (CVS) for subordinate units into command and staff training exercises would thus defeat the purpose of these exercises and the provision of CVS for the commander and his staff is envisioned as providing only an extremely marginal payoff since they could only be used to view terrain, not the actual unfolding of the battle. Second, if you should add CVSs to command and staff exercises,

you would in reality have a SIMNET supported collective training exercise, such as a Field Training Exercise or a Tactical Engagement Simulation Exercise. The preliminary conclusions reached after this type of analysis indicates that the singularly highest payoff from SIMNET will come from its replacement of collective training exercises within field units and as such this portion of the research program has focused SIMNET's use in that capacity.

The second factor which impacts on an investigation of how SIMNET can best be utilized is a function of the training management and training process within a unit. This process has recently been redefined in Field Manual 25-100, Training the Force, and will be further articulated in the soon to be completed Field Manual 25-XX, Training the Battalion. The overall process from which the training management cycle in FM 25-100 is derived is depicted in Figure 3.

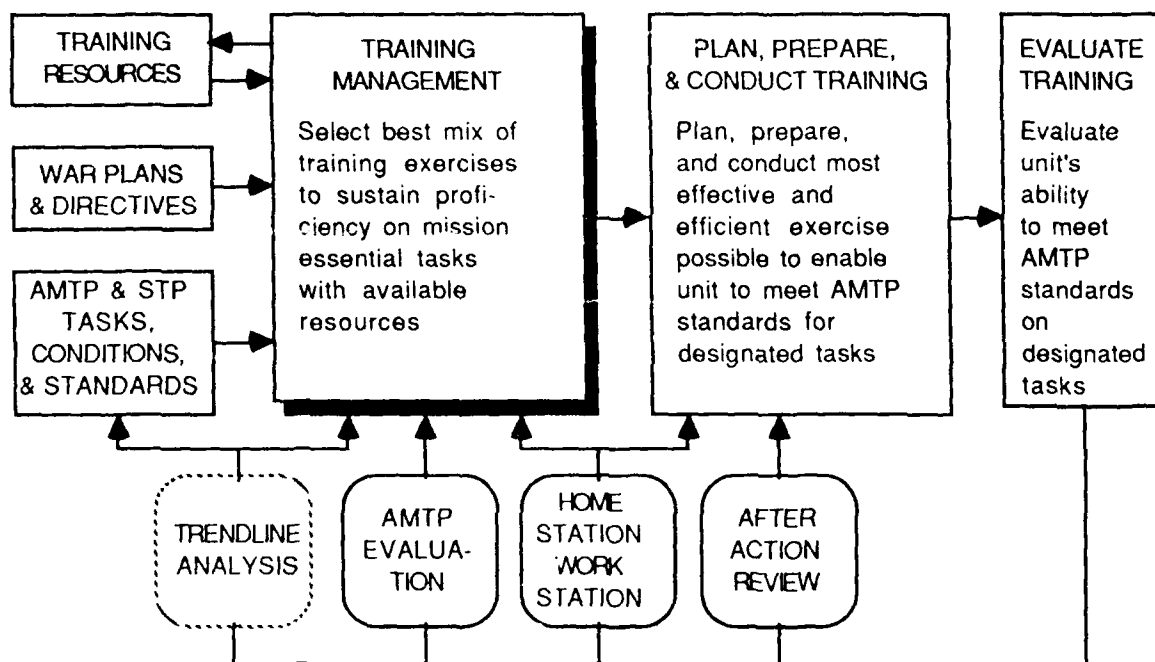


Figure 3 - Unit Training Process

The research effort defined by the subtasks listed at the beginning of this paper focuses on the training management component of this process. Other subtasks within the overall ARI research effort focus on various other components such as the evaluation criteria and the planning and preparation of a training exercise. The relationship of the Home Station Workstation concept to SIMNET and the applicability of the Army's Trendline Analysis program of SIMNET data, however, are beyond the current scope of the ARI program.

An expansion of the training management component is depicted in Figure 4.

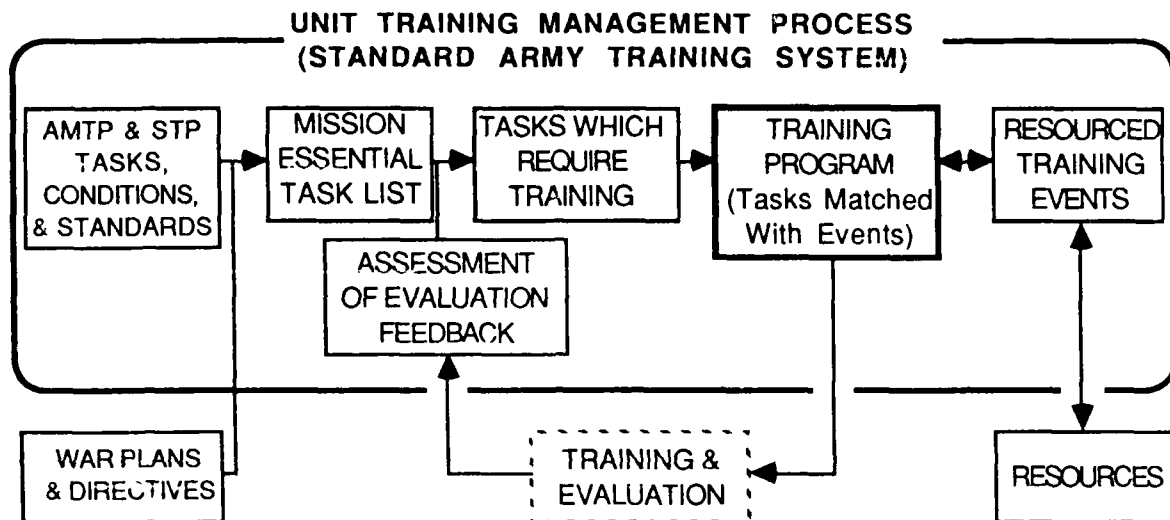


Figure 4 - Unit Training Management Process

As indicated, this entire process culminates with the development of a training program or plan which matches Mission Essential Task List (METL) tasks which require training with resourced events such as a SIMNET supported collective training exercise.

Automation support to assist unit commanders implement this process is currently under development within the Standard Army Training System (SATS) program (formerly the Integrated Training Management System program). As such, no additional SIMNET unique automated training management support is required for SIMNET. This does not preclude the necessity, however, for automation support for the planning, preparation, conduct, and evaluation of a SIMNET exercise once it has been selected within the management process and the requirement for such automation is currently being examined within other tasks within the overall ARI program.

The real issue at hand is thus the match between tasks which have been selected for training within SATS and the training events (or exercises) which can be resourced. If SIMNET could support all tasks, this would be a uninteresting problem. Initial indications, however, indicated that SIMNET in its current form cannot support all of the tasks contained in current ARTEP Mission Training Plans (AMTP), which in turn will form the basis for the derivation of most unit's METLs.

The primary focus of the research program described in this paper is the determination of the relationship of AMTP tasks to alternative training exercise options.

Certain tasks, such as "Breach Defended Obstacle," "Perform Air Assault," and "Perform NBC Operations" are clearly beyond the capability of the current SIMNET system to support. Other tasks, such as those indicated in Figure 5 require a more detailed analysis at the subtask level.

TF ASSAULTS	SIM	FTX	BS	CFX	ST	TEWT
SCOUTS DETECT THE ENEMY	H	H	M	M	M	
TF DEVELOPS THE ENEMY SITUATION	H	H	H	H	M	M
TF DEVELOPS AND COMMUNICATES A FRAGO	H	H	H	H	M	
TF FIXES THE ENEMY	H	H	H	H	M	

TF DEFENDS

TF PLANS THE DEFENSE AND ISSUES OPORD	H	H	H	H	M	M
BATTALION TF PREPARES THE DEFENSE		H		M	M	H
TF DEFEATS THE ENEMY RECON & COND SURV	H	H	H	H	M	
TF DEFEATS THE ATTACK	M	H	H	H	M	

TF PERFORMS RELIEF IN PLACE (NIGHT)

TF PLANS RELIEF IN PLACE	H	H	H	H	M	M
TF IS PREPARED FOR RELIEF IN PLACE		H		M	M	
TF CONDUCTS THE RELIEF		H		M	M	
TF REACTS TO ENEMY ATTACK		H	H	H	M	

Figure 5 - Analysis of Match Between Subtasks and Alternative Exercises

The coding represents either a high order match (H), a moderate match (M) or no match. As indicated, little problem is encountered with the offensive task of "TF Assaults." With regard to the defensive task of "TF Defends," however, serious limitations exist with regard to the capability to support the subtask of "Battalion TF Prepares a Defense" -- vehicles cannot be dug in or camouflaged, routes cannot be marked, personnel cannot be dismounted and placed in observation posts, etc. As such, if the SIMNET is utilized to support this task, the unit's training program should schedule a Tactical Exercise Without Troops or some other appropriate exercise to augment the SIMNET exercise with respect to this subtask. An finally, the figure indicates that the subtask analysis of the task "TF Performs Relief in Place (Night)" cannot effectively be supported with a SIMNET exercise.

Once these task to event relationships have been identified (in a far more rigorous manner than indicated here) they will be incorporated into SATS and also very likely serve as the basis for prioritization of further SIMNET upgrades and related training system developments.

THE SIMNET RESEARCH PROGRAM: AN OVERVIEW

Thomas J. Lubaczewski
Perceptronics, Inc.
2500 South 4th Street
Leavenworth, KS 66048

ABSTRACT:

This paper provides an overview of the research objectives of the current effort. Although the effort is an exploratory research effort, the research team is committed to conducting the research with a conscious intent to insure that the research is directed toward the training requirements as perceived by the leaders in the Army who are charged with training development activities.

PROGRAM OVERVIEW

The SIMNET research effort entitled Unit Performance Measurement and Training Program Design for Networked Simulators has as its principle goals, the development of a performance measurement system that is common to both the National Training Center (NTC) at Fort Irwin and the SIMNET training system currently being introduced within the Army (including the application of SIMNET technology to the Close Combat Tactical Training (CCTT) System, the production version of the prototypical SIMNET that the Army plans to procure in the near future); the development of a training strategy that will facilitate the employment of SIMNET in such a way that its benefits are maximized; and finally, the development of an overall strategy that investigates SIMNET (and the CCTT) into the full array of training devices and events that are available within the total Army training system.

As a start point, the development of a common performance measure system will key on the relevant ARTEPs, AMTPs and the 2A training objectives being developed to support the training and evaluation of units at the National Training Center. Based on these objectives, a conceptual performance measurement system that is common to both the NTC and SIMNET will be developed. Automated performance measurement, especially in the SIMNET environment, will be stressed to preclude requirements for a large number evaluators/controllers. Assessments of how well units perform their missions and tasks, necessary for the conduct of this research, will at all times be handled on a strictly

confidential basis. In addition, the results from use of the performance measurement system will be subjected to scrutiny by subject matter experts from within the Army training community to insure that the impact on random variables on mission outcomes do not unduly influence the assessments of a unit's performance. The benefit of using SIMNET to evaluate a unit, especially prior to its NTC rotation, is that observation of a unit during a simulated battle is possible on a completely non-intrusive basis. Moreover, unit leaders can fine-tune their skills, and those of the other members of the unit prior to arrival at the NTC, thereby making the time spent at the NTC more beneficial.

How best to use SIMNET is of paramount importance and will be addressed in this effort. There are those who maintain that a SIMNET exercise should remain strictly free-play with no controller input whatsoever. Others maintain that since time available to train on SIMNET is a resource that is constrained (just as ammunition and fuel are constrained) and that it is better to have some degree of control so that maximum learning can take place. When, how and to what degree this control should occur will be explored. Another element that will be considered is whether certain prerequisite individual and collective skills, and to what degree, should be mandated before a unit is scheduled to utilize SIMNET. Another area that will be explored is whether there are some key unit indicators, as identified in an organization's mission essential task list (METL), that provide a reasonably high level of reliability in predicting success in mission accomplishment.

Finally, there is the question of how to fit SIMNET (and the follow-on CCTT) training into the variety of devices, training events and decision aids currently, or soon to be, available to the Army's commanders and training managers. This overall strategy has often been overlooked in the past. If the Army is to realize the full advantage of the technology represent in SIMNET, it must have researched based information on which to develop such a strategy. Because SIMNET provides a unique training experience and is intended to be distributed to a broad-scale, it is essential that it be integrated with such things such as live-fire exercises, TEWTs, FTXs, CPXs, UCOFT training and command group battle simulations. Sequencing of SIMNET training in a unit's training schedule, amount of time spent on SIMNET training, and the use of SIMNET for remediation are just a few of the critical issues that will be considered. Along this line, since the Army appears to be ready to adopt the Integrated Training Management System (ITMS) as its method for managing unit training, it is important that this effort explore the relationships that

must be developed to insure that this SIMNET research effort and ITMS are consistent in their goals.

FOCUSED OBJECTIVES

In order to provide a more focused view of what this effort seeks to accomplish, the following list of research objectives that have been provided to the research team by the Army Research Institute for the Behavioral and Social Sciences (ARI). They capture the essence of the goals that will guide the effort of the research team throughout the performance on this effort.

- Design a performance measurement system for the SIMNET technology keying off the ARI/CATA research on an improved ARTEP that has focused on the NTC for development.

- Design and evaluate a training management system for the SIMNET technology; examples of issues are:

- Identification of the prerequisite skills that ensure positive transfer to operational equipment and avoid negative, dangerous transfer, such as driver safety errors;

- Design of the kinds of scenarios that should be used at particular points in the learning curve;

- The kind of trainer feedback to be given, and when and how; for example, the decision rule for stopping an exercise and deciding what to do next.

- Design training strategies for incorporating SIMNET training into unit programs along with UCOT, the Battle/Brigade Simulation (BBS), FTX, MAPEX, CPX, NTC, etc.

BENEFITS OF RESEARCH

FM25-100 (Draft) states that leaders must train each individual within the unit to perform his job in war as a key member of a capable unit. This training is to be based on wartime requirements and must guide the planning, execution and assessment of each organization's training. Organizations cannot be proficient in every possible training task; therefore, commanders must use a battle focus process to selectively identify those training tasks essential to accomplishing the organization's wartime

mission. This focus is derived from compiling applicable tasks, as identified in external directives, and selecting for training only those tasks which are essential to accomplish their organization's wartime mission. The compilation of these tasks is the organization's mission essential task list (METL). NTC, SIMNET, CPXs, FTXs, and command and control battle simulations are a few different ways of conducting training to achieve/maintain combat proficiency in the organization's METL. Successful performance at the NTC or in training conducted on SIMNET is a valid indicator to wartime preparation only if the tasks included in SIMNET and NTC correspond to the unit's METL and the missions exercised reflect wartime conditions.

The results of this research effort will investigate Army goals for home station, NTC and SIMNET training which support the Army's doctrine for training the force. In accomplishing this, the research team will develop/identify:

- Techniques for measuring unit performance;
- Validation procedures for SIMNET training;
- Methods for determining the value of procedural task performance measurement;
- Differences, and their impact, between NTC and combat; and
- A global training strategy for SIMNET.

As progress is made, reviews will be conducted with Army training development activities to insure that Army concerns are addressed during the research effort. The philosophy that the research team is committed to conducting the research not for the sake of research alone but rather to lead to the development of useful products that will address the concerns and requirements of the Army.

These research areas, and the results, will be directed toward providing the Army with a framework from which decisions can be made and procedures implemented which provide, the Army force, the most productive and beneficial training necessary, as stated in FM 25-100 (Draft), to: "develop and maintain warfighting skills for the Army to fulfill its fundamental mission: deter war, or, if deterrence fails, reestablish peace through victory on the battlefield."

STRATEGIES FOR UNIT PERFORMANCE MEASUREMENT IN SIMNET

Nancy K. Atwood & William J. Doherty
The BDM Corporation

Simulated Networking (SIMNET) is currently one of the focal points of attention within the Army training community. As described by other papers in this symposium, SIMNET offers the potential for powerful and cost effective collective training on a simulated battlefield. One important component of a SIMNET training program is a measurement system for assessing unit performance. The purpose of this paper is to describe research on a methodology for developing a system for measuring the performance of units training with SIMNET.

The Requirement for Unit Performance Measurement

FM 25-100, the Army's capstone training manual, lays out the Army's overall training philosophy and establishes the doctrine for Army training management. The concept of train-evaluate-train stands at the heart of this philosophy. Training is seen as an iterative process which requires explicit statement of training objectives and requirements, formulation and execution of a training strategy, and an evaluation of performance.

Performance assessment is a cornerstone of the training management cycle. A system for measuring performance is required to assess a unit's capability in order to determine training requirements. Such a system is also required to assess the outcomes of training and to determine needs for follow-on training. In essence, a performance assessment system provides the diagnostic feedback that ties the entire training management cycle together and allows leaders to assess the strengths and weaknesses of their unit (i.e., their training status).

SIMNET's capability to provide combined arms collective training makes performance measurement particularly important. This capability offers the potential for providing rigorous and stressful training to units focused on advanced training objectives similar to those of the Army's Combat Training Centers (CTCs) such as the National Training Center at a fraction of the cost. However, to approach the training benefits of the CTCs, SIMNET must satisfy CTC objectives to provide performance feedback to training units (described in AR 350-50 draft). Thus, design of a SIMNET training strategy must include a system for measuring unit performance in order to meet the requirements of Army training doctrine and to provide a platform for SIMNET to serve as a full fledged contributor to the Army's of advanced collective training program.

Objectives of the Research

The objective of this component of the SIMNET research program is to design a methodology for developing a system for measuring unit performance in the SIMNET environment. More specifically, the goal is two-fold: 1) to develop a measurement system for SIMNET that maximizes commonalities (as appropriate) with the Army's most realistic heavy combat simulation, the National Training Center; and 2) to capitalize on the unique performance assessment capabilities inherent in SIMNET.

The above goals were formulated to ensure that the system developed for assessing units training with SIMNET has the greatest diagnostic power possible. First, by maximizing commonalities with the NTC, heavy units can adopt an integrated training strategy using both computer simulation (SIMNET) and tactical engagement simulation (NTC) as components of their training program where unit performance can be assessed with common benchmarks. Second, by capitalizing on the unique capabilities of SIMNET, more specific feedback on performance is possible than would normally be available in a field training environment such as the NTC.

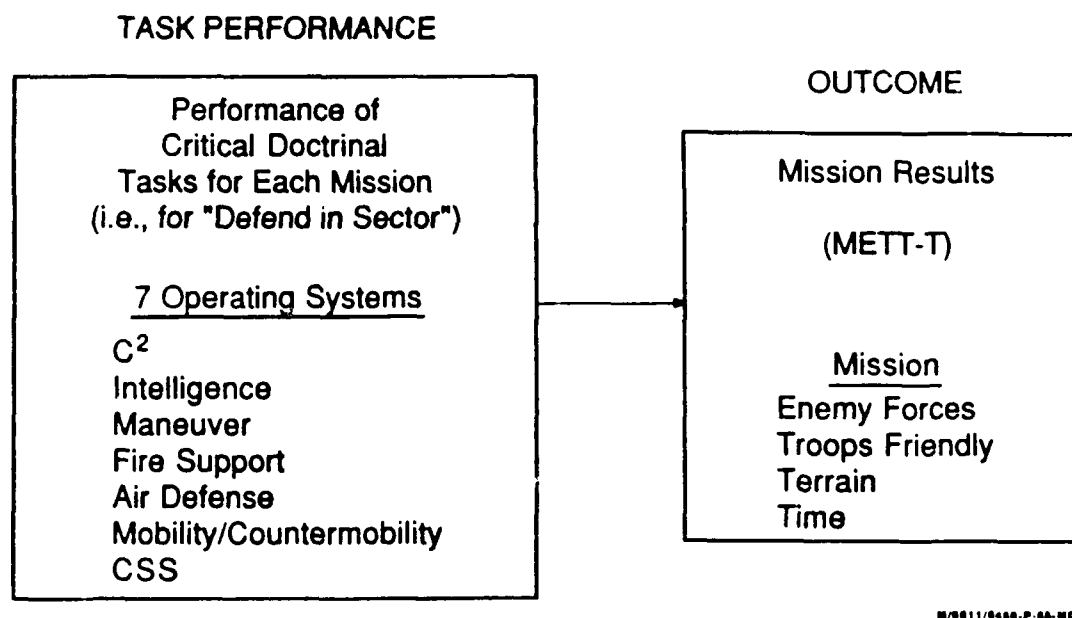
Approach for Unit Performance Measurement

The obvious approach to assessing unit performance in SIMNET is to use the Army Training and Evaluation Program (ARTEP) and the recent ARTEP Mission Training Plan (AMTP) Program. Doctrinally, the Training and Evaluation Outlines (T&EOs) included in the AMTP are intended to serve as the foundation for measuring unit performance. As such, they include an identification of training tasks along with a statement of performance conditions and standards.

While the ARTEP and AMTP programs provide the foundation for performance assessment, they are insufficient in their current form for direct application in a complex collective training environment such as the NTC or SIMNET. The lists of training tasks are too long and lack prioritization or an organizing structure for observers to use reliably. Further, because the ARTEP/AMTP programs were designed for use in home station training, they do not take into account the unique capabilities of computer driven simulations for measuring unit performance. Finally, they do not include a strategy for assessing overall mission effectiveness as a benchmark for examining task performance.

The Army Research Unit Field Unit at the Presidio of Monterey has undertaken a research program to develop methodologies to strengthen the AMTP as a unit performance measurement tool (see Levman, 1987; Root and Zimmerman, 1988). The focus of this research has been the NTC, a natural counterpart for SIMNET because of its orientation on heavy units. Our approach in this project is to capitalize on this research and to use it as a starting point.

The ARI-POM unit performance methodology is based on a two-part model as shown in Figure 1. The model recognizes the role of both process and outcome. The mission results portion of the model provides a "bottom line" outcome measure of the degree of mission accomplishment. The critical task performance component provides an immediate process measure of unit behaviors that contribute to mission accomplishment.



SI/0011/0400-P-00-00

Figure 1. Unit Performance Measurement System Model

In applying this model to SIMNET, our approach has been to examine the applicability of the NTC-based model to SIMNET and to identify the boundaries of overlap as well as the unique measurement capabilities of SIMNET for both components of the model. In developing measures in each of these areas, we have given priority to those that can be collected from the automated data stream in an effort to increase reliability and reduce the data collection burden on observers. The remainder of this paper provides examples of outcome and process measures and concludes with key issues being encountered.

Measures of Mission Effectiveness

As shown in Figure 1, the mission accomplishment indices in the outcome portion of the model are structured around the accepted military analysis factors of METT-T (mission, enemy resources, friendly troops, terrain and time). These indices taken individually provide useful descriptors of aspects of mission accomplishment. When combined, they provide a benchmark of mission success or failure.

Analysis of the applicability of mission accomplishment indices to the SIMNET environment indicated that the NTC-based indices were largely applicable. For both defensive and offensive missions, measures of enemy attrition, friendly attrition, and terrain control were judged appropriate and easily gathered from the automated data stream supporting the simulation. Measures of time to complete the mission, while easily collected, are heavily influenced by the scenario. Thus, time indices are currently receiving only tentative consideration until a more thorough analysis of the range of likely scenarios can be completed.

In summary, it is important to note that mission accomplishment indices are only meaningful at the echelon level to which mission has been assigned. For example, at the NTC missions are most often battalion task force missions; thus mission accomplishment is examined at the battalion level. Computing similar indices at the company or platoon level is not meaningful because of the interdependent contribution of the component echelons to the accomplishment of the larger battalion level. The same reasoning applies within SIMNET; mission accomplishment must be assessed at the echelon level to which the mission has been assigned.

Measures of Critical Task Performance

The diagnostic power of the measurement system lies in the measures of critical task performance. It is the performance of doctrinally based tasks, validated by military experts, that provides the explanations for why their missions succeeded or failed.

The methodology for assessing critical task performance developed at ARI-POM has several important features. First, it is based on a collective front end analysis that draws on doctrinally identified tasks. As part of this analysis, "critical" tasks were culled based on extensive input from military experts. Second, the tasks have been organized into a battle flow framework which identifies battle phases and segments with a specific purpose and observable outcome. The critical tasks have been organized within this framework and are linked within this structure to show sequential and functional linkages.

Our analysis of the applicability of tasks within the SIMNET environment was conducted by categorizing tasks into one of the following categories according to whether it could be: performed in SIMNET device, performed in the SIMNET complex, performed in the SIMNET complex with modifications or additions, or not supported.

The results of this analysis indicated that with minor exceptions all planning and preparation tasks could be performed in the SIMNET complex as is or with minor additions and modifications. In addition, most of the execution tasks could be performed in the SIMNET device with some

exceptions (such as executing an obstacle plan, responding to NBC operations, conducting evacuation procedures). In summary, the overlap in critical task performance at NTC and in SIMNET is considerable. These parallels are encouraging because they support unit training programs which are integrated and mutually reinforcing.

Issues In the Development Process

We are currently working to operationalize the measures of performance for critical tasks within the SIMNET environment. We expect that assessment of planning and preparation tasks will require judgment by observers as at the NTC. However, many execution tasks--particularly those involving movement and firing--can be assessed using data from the automated data stream. Furthermore, the SIMNET data stream allows for more detailed measures of performance (such a turret orientation) which cannot be gathered in a field setting such as NTC. Finally, automated measures of the performance of execution tasks will need to be supplemented by military experts monitoring unit performance using the plan-view display and monitoring the communications net. It is only through the integration of multiple data sources that a comprehensive picture of unit performance can be achieved.

References

Lewman, T.J. (1987) Mission critical tasks at the National Training Center. (ARI Research Product) Alexandria, VA: U.S. Army Research Institute.

Root, J. & Zimmerman, R.A. (1988) A unit effectiveness measurement system at the National Training Center. (ARI Research Report) Alexandria, VA: U.S. Army Research Institute.

Feedback Strategies for SIMNET

Judith J. Nichols and James W. Kerins
The BDM Corporation

The Army-wide implementation of SIMNET will provide units with a relatively low cost mechanism for routinely practicing the collective skills inherent at battalion, company, platoon, section, and squad/crew echelons and the command and control staff actions associated with multi-echelon combat scenarios. SIMNET's focus is the close combat heavy battle.

This paper examines the various feedback mechanisms available to SIMNET trainers -- e.g., the After Action Review (AAR) -- and describes a strategy for feedback delivery which encompasses both current and developing Army feedback capability.

SIMNET Feedback Strategy

The presentation of diagnostic feedback that enables units to identify their strengths and weaknesses and plan follow-on training plays a critical role in the Army's train-evaluate-train philosophy. The SIMNET feedback strategy is designed to provide a system which:

- (1) Utilizes current Combat Training Center (CTC) state-of-the-art technology;
- (2) Exploits SIMNET replay and visual presentation capabilities;
- (3) Identifies feedback intervention points which maximize opportunities for corrective and/or sustainment training guidance;
- (4) Identifies the most appropriate type of feedback for each intervention point;
- (5) Capitalizes on a measurement system that maximizes commonalities with the National Training Center (NTC); and
- (6) Facilitates follow-on training by providing units with accessibility to their SIMNET unit performance data and assessments through use of the Integrated Training Management System (ITMS) at the units' home stations.

CTC Feedback Technology

The Army currently has four Combat Training Centers (CTCs) in various stages of operation: the National Training Center (NTC), Fort Irwin, CA, the Battle Command Training Program (BCTP), Fort Leavenworth, KS, the Combat Maneuver Training Center (CMTC) at Hohenfels, FRG, and

the Joint Readiness Training Center (JRTC), Fort Chaffee, Arkansas. Although each CTC has a different training focus, all CTCs share the same train-evaluate-train philosophy and feedback mechanisms (i.e., After Action Reviews and a Take Home Package -- THP).

CTC training is the most advanced collective training currently available to Army units. The demands placed upon the CTC unit performance evaluation and feedback systems has produced a state-of-the-art feedback technology which can contribute greatly to the SIMNET feedback process. For example, the NTC feedback environment combines digital data collected by the NTC Instrumentation System, both formal and informal observations made in the field, exercise replay capability, field video, tactical communications, statistical graphics capability, and training analysis expertise to produce one of the most sophisticated AARs available in the Army today.

In addition to real-time feedback (i.e., AARs), CTC feedback technology has also produced the Take Home Package, a post exercise archive of unit training. A SIMNET THP would facilitate unit follow-on training by providing a record of unit SIMNET performance which included mission summaries, identified unit strengths and weaknesses, and other pertinent data (e.g., personnel and equipment casualties, number of rounds fired, etc.).

Although the SIMNET environment is not scheduled to be resourced with either the personnel or the technical support system available at the NTC and other CTCs, it does have the potential to produce many of the same AAR support aids as the NTC (e.g., exercise replay capability, automated data collection, O/C data collection and analysis) and a SIMNET THP.

SIMNET Replay and Visual Representation Capabilities

The SIMNET systems' data stream provides a continuous update of information on such items as vehicle/weapon system status, locations, vehicle or weapon system orientation, and firing events in a manner very similar to that of the NTC Instrumentation System. The information, routinely gathered by the system, represents an untapped source of critical training feedback data that can be fed back and utilized by unit leaders to assess the present state of a unit's training and to identify future training needs.

The Plan View display, presently available in the SIMNET system, provides an immediate capability to playback a unit's entire training mission or activity. This display provides unit leaders with the ability to observe and critique the unit's actions and performance from start to finish. The ability to play the entire mission or training activity at hyper speed, focus on a critical segment, or stop the action to examine a critical event gives the unit commander a powerful tool with which to examine, analyze, and critique his unit's performance. The Plan View display becomes the central element of the SIMNET AAR.

In conjunction with the Plan View display the routine information captured by the SIMNET data stream can be collated and displayed graphically to further aid the unit commander in analyzing his unit's performance. The data, displayed as graphs and charts, would provide such information as the number fired in a given time period, the type of ammunition fired, the number of hits achieved, the number of vehicles or weapon systems destroyed or "killed", the ranges at which hits and "kills" were achieved, etc. This information, organized and displayed in simple and easy to read graphs and charts, provides the unit commander with valuable insights into his units' performance. Further, the information and data will assist the commander in determining the training required to improve the unit's performance.

In addition to the graphically presented statistical data the unit commander also has available data and information that can be portrayed or displayed on the Plan View display. This data includes such items as where friendly indirect fires were actually placed or delivered, the location of damaged or destroyed vehicles and weapon systems (both friendly and enemy), the location of friendly and enemy obstacles, the location of impact of enemy indirect fires, etc. As with the statistical data, this information provides the unit commander with important information and insights into his units' actions and performance. Additionally, this data, as with the statistical data, will assist in determining the training needs of the unit.

Feedback Intervention Strategy

Once the potential for SIMNET feedback mechanisms (AARs and THP) has been defined, it will be necessary to identify feedback intervention points which maximize opportunities for corrective and/or sustainment training guidance. While intervention for the purposes of feedback should remain flexible enough to accommodate unit trainers' specific training procedures and requirements, the overall feedback strategy must identify those points where routine feedback will maximize training value. For example, AARs are provided at the close of each mission at the NTC. This particular intervention point has demonstrated its training value repeatedly. The SIMNET feedback intervention strategy should also include guidance in connection with such occurrences as discontinued missions, repeated missions, premature failure on the part of the unit, SIMNET equipment failure, etc. Formal and informal AARs, coaching, repetitive training, and other feedback mechanisms must be included and identified as the preferred or alternative method of feedback at the various training intervention points.

SIMNET Follow-On Training

In addition to real-time training and feedback, the SIMNET feedback strategy also ensures and facilitates quality follow-on training by providing units with accessibility to their SIMNET unit performance data and assessments through use of the Integrated Training Management System

(ITMS) at the units' home stations. SIMNET allows a commander to tailor the training to exactly match the unit's skill level and training needs. Utilizing information from ITMS a commander can structure a SIMNET training exercise that is specifically designed to correct specific training deficiencies or weaknesses. At the conclusion of the SIMNET exercise the commander has available to him information and data that allows him to immediately update the unit's training status within ITMS and begin his planning process for future training. The integration of SIMNET training feedback with ITMS will provide the unit commander with an efficient means of identifying the most economical training strategies for correcting training deficiencies and sustaining and maintaining unit training proficiency.

Summary

The implementation and fielding of SIMNET will provide units with a low cost mechanism for practicing the collective skills required at the battalion/task force echelon through the squad/crew level and the command and staff actions associated with multi-echelon combat scenarios.

The presentation of training feedback that enables units to identify their strengths and weaknesses and plan follow-on training plays a critical role in the Army's train-evaluate-train philosophy. SIMNET's feedback strategy is designed to provide a system that:

- (1) Exploits SIMNET replay and visual presentation capabilities;
- (2) Identifies feedback intervention points which maximize opportunities for corrective and/or sustainment training guidance;
- (3) Identifies the most appropriate type of feedback for each intervention point;
- (4) Utilizes a measurement system that maximizes commonalities with the National Training Center (NTC); and
- (5) Facilitates follow-on training by providing units with accessibility to their SIMNET unit performance data and assessments through use of the Integrated Training Management System (ITMS) at the units' home stations.

SIMNET's training feedback strategy must recognize and incorporate the fact that performance assessment and diagnostic feedback are critical to successful training. A performance measurement system is an essential requirement in determining a unit's capabilities and its future training requirements. Performance assessment provides the diagnostic feedback that links a unit commander's assessment of his units strengths and weaknesses to the training management process of determining the requirements for follow-on training.

ISSUES IN MEASURING UNIT PERFORMANCE AT COMBAT TRAINING CENTERS

William J. Doherty, Chair
The BDM Corporation
Monterey, California

This panel examines critical issues in assessing the performance of units training at the Army's Combat Training Centers (CTCs). Reliable and valid measurement is vital to ensuring full training benefits of advanced collective training, providing targeted diagnostic feedback to training units, and developing performance data for identifying systemic lessons learned for the Army at large. Dr. Jack Hiller introduces the criterion problem of performance measurement, identifies obstacles to measuring combat effectiveness, suggests methodological approaches for solving the criterion problem, and highlights the practical limitations of unit performance data. Subsequent papers explore key features of the training environments of the Army's three currently operational CTCs and examine their implications for measuring unit performance. Mr. Root presents an assessment system developed for use with heavy Battalion task forces at the National Training Center (NTC) and highlights issues encountered in extending the system to brigade level. Ms Nichols and LTC Crawford describe approaches to assessing performance of light units and special forces training at the Joint Readiness Training Center (JRTC). Mr. Ross and Dr. Girdler analyze the requirements for assessing performance at the Division and Corps levels at the Battle Command Training Program (BCTP). Finally, Drs. Doherty and Atwood examine the role of Simulated Networking Technology (SIMNET) as a surrogate in advanced collective training and performance assessment. The program concluded with the discussant, MAJ McLaughlin reflecting on issues raised from the behavioral measurement and combined arms training perspectives.

PRACTICAL SOLUTIONS TO THE CRITERION PROBLEM AT THE COMBAT TRAINING CENTER

JACK H. HILLER

U.S. ARMY RESEARCH INSTITUTE

Introduction

The criterion problem facing the Army at its combat training centers is not really a single problem, but rather, many opportunities. Measures of unit performance, collected while units train under conditions of realistically simulated combat at the centers, offer potential insights for improving a variety of systems:

1. The Performance of Collective Training Centers
 - a. Management and conduct of training at the centers
 - b. Preparation of take-home training feedback
 - c. Support of research, e.g., data collection
2. Training and Operations at Home Station
 - a. Management and conduct of training
 - b. Leadership principles and practices
 - c. Policies and practices affecting unit cohesion
3. The Validity of Tactical Doctrine
 - a. Command, control, and intelligence
 - b. Movement, maneuver and direct fire
 - c. Combat support and combat service support

In fact, virtually every aspect of Army policy, doctrine, and operations may potentially be examined, using measured performance as the criterion of effectiveness. Clearly, the performance of units who have prepared for combat may reveal strengths and weaknesses in our systems--but performance is hard to measure. In this paper I will review several basic problems associated with unit performance measurement and describe a major research project now underway.

Performance Measurement Problems and Solutions

PURPOSE. One fundamental measurement problem precedes the act of measurement and concerns, instead, the need to clarify and be specific about the intended use or purpose of the measures. During the early days at the National Training Center (NTC), Ft. Irwin, California, while it was being organized and built with difficulty, the commander was reported to have said that researchers should stay away, since he would see to it that all of the data were collected and shipped, as if data form a discrete, definite entity. I have to admit that researchers did not at that time know what data should be collected for what purpose. In contrast, a recent NTC commander explained to me, as I started to address the topic of data collection, that the data

routinely collected (mainly digital data describing the positions and firing events for major weapons systems) never had the focus or precision required to enable an understanding of what exactly happened and why, or how unit performance could have been improved. He thereupon volunteered to collect data specifically required for targeted research issues. In recognition of this insight, specific issues are now selected for study during NTC training missions (for example, Commander Survivability on the Battlefield); and the observations to be made are planned well in advance and performed by subject matter experts.

TRAINING CONFORMITY TO TACTICAL DOCTRINE. Tactical doctrine and training literature provide basic guidance for training at home station and at the combat training centers alike. A key issue concerns how well the units execute doctrine, that is, how closely their performance conforms to doctrine. Since there are hundreds of tasks described by the training guidance literature [Army Training and Evaluation Programs (ARTEPs) and Army Mission Training Plans (AMTPs)], it was not originally thought to be practical to attempt to observe, score, and report performance on these many tasks. At the NTC the original strategy for generating training feedback was to focus on the several operating systems, e.g., maneuver, fire support, command and control, intelligence, air defense, etc. This strategy proved to be manageable but also had problems. Since the unit training programs are directly based on missions and tasks as organized in ARTEPs/AMTPs and not operating system descriptions, the large take-home packages given to units by the NTC were therefore not seen as directly useful for planning and conducting home station training.

The data, based on the unstructured observation of operating system performance by NTC Observers/Controllers, also had limited utility for research. At one point a few years ago, the Vice Chief of Staff directed ARI to analyze NTC data to determine performance trends. Since the take-home package data reflected self-selected, subjective observations by the NTC Observer/Controllers, it was impossible to analyze performance trends. It was also difficult to analyze take-home package data to reveal deficiencies in specific mission (or ARTEP) tasks.

To overcome the inherent difficulty of using the many detailed ARTEP/AMTP task definitions, ARI has pursued two research approaches. The training literature was intensively analyzed by ARI and BDM for completeness (deficiencies were found in representation of planning and preparation tasks), and then all candidate tasks were reviewed and rated for criticality by experts at the NTC, the Combined Arms Training Activity (research sponsor), and the Infantry and Armor schools. The other approach was to invent an "Electronic Clipboard" to simplify and automate the functions of:

- * recalling and presenting to the observer checklists for scoring performance;
- * recording and storing task scores by unit, time, and mission;
- * uploading the stored scores to a computer to enable rapid analysis and training feedback, as well as longterm use for research.

A portable, handheld device was developed by ARI, the Jet Propulsion Laboratory and PERCEPTRONICS, and successfully demonstrated. The status of work to implement the Electronic Clipboard concept at the NTC and Joint Readiness Training Center (JRTC) is described by Patrick Whitmarsh at this conference. We may also be optimistic that the concept will be applied to performance measurement and analysis in the Battle Command Training Program.

VALIDITY OF TACTICAL DOCTRINE. The training guidance literature is based on a multitude of analyses that range from elements of basic psychology (surprise, deception, fear, esprit), to rudimentary job-task analyses for operating weapons systems, to tactical and strategic planning, to complex command and control of combined arms and joint service air power. These analyses are based on past and recent histories, weapons systems performance demonstrations in bench tests (maximum performance capabilities), mathematical models, subjective feedback from the field on draft versions of training and doctrine literature, and a whole lot of intuition. To a modest extent, intuition has been guided by training experiences at home station. The Combat Training Centers obviously provide a rich source of experience for examining the adequacy of current training and doctrine. The caveat is that subjective evaluation of performance is notorious for low reliability--experts will disagree on what happened as well as why.

To discipline the use of expertise, it is useful to have objective measures of performance outcomes as a check on validity. Fortunately, the NTC routinely generates objective measures in the form of position/location and hit/kill data based on a system of lasers simulating munitions (MILES) and radio telemetry. There are a variety of cautions that need to be considered when using the NTC MILES data (see Hiller, 1987), but such objective outcome measures are indispensable for gauging the validity of training and doctrine.

A MAJOR RESEARCH PROJECT: RESOURCES TO READINESS

The level of training resources required to sustain training readiness in the heavy ground forces is currently regarded as a serious question by OSD and the Congress. The Army has answered

this question in the past based on expert estimates of training activities required and the readiness reports prepared by field commanders. This section of the paper: a. identifies weaknesses inherent in training readiness assessments prepared at home station; b. addresses the potential use of the NTC for generating objective information and, c. explains an analytic model for relating training resources to readiness.

HOME STATION TRAINING READINESS ASSESSMENTS. The unit commander is required monthly to estimate how many weeks it would take him to bring his unit to a complete state of combat readiness, given that he had available all of the training resources necessary. To form this estimate, he must know the actual state of training readiness. How does he acquire this knowledge? He reviews individual and collective training records and his own personal recollection of unit performance during training activities. Unfortunately, most unit home station training does not provide combat realism because of limitations in the availability of terrain that resembles the unit's war time mission assignments. Likewise, the opposing forces available to create realism vary from non existent, to poorly equipped, to excellent (at only one home-station post). The unit commander is thus forced to make a complicated subjective estimate of readiness.

NTC TRAINING: A REALISTIC COMBAT SIMULATION. Although the NTC does not now provide a complete combat simulation--because of deficiencies in instrumentation for indirect fire, mobility and counter mobility, air and air defense--it does stress heavy forces in the use of their organic weapons systems. If a question to be answered concerns the minimum resources required for training armor and mechanized infantry task forces, then the performance of these units while they train at the NTC provides a focused criterion. There are three specific forms of measurement that have been proposed (Hiller, 1987) to assess a unit's performance capability:

a. Conformity to Tactical Doctrine (task performance scored according to standards specified in ARTEPs/AMTPs);

b. Mission Outcomes, as measured by an index formed from three variables:

- * Percentage of friendly forces (major instrumented weapons systems) remaining at the end of a combat mission, as measured by the MILES;

- * Percentage of enemy forces killed;

- * Terrain Control, as measured by the percentage of weapons systems crossing a defensive boundary.

Alternatively, a traditional casualty exchange ratio may also be calculated.

c. Expert Judgement. The primary need for expert judgement is to gain creative insights on battle performance that may be missed by mechanically scored outcomes and task performance.

Expert judgement is also required to decide if any given mission is representative of the mission played or if unusual circumstances render data unusable or especially interesting. Expert judgement may also be useful in rating overall performance on key dimensions or categories such as Move, Shoot, and Communicate.

These three forms of criterion measurement, i. e., task performance, mission outcomes, and expert ratings of performance, may be used as independent performance criteria or combined to form a single criterion measure.

TRAINING RESOURCES AND OTHER PREDICTORS. The major training resources of interest concern vehicle mileage for tanks and carriers, and gunnery training (live fire) expended in the primary NTC train-up period, approximately six months. However, relatively inexpensive forms of simulator based training (given that the simulators are now in the inventory) may be predictive of performance; these include the Unit Conduct of Fire Trainer (UCOFT), and the ARTBASS command control trainer. Potentially significant moderator variables include a rating for similarity of home-station training to NTC, and a measure of personnel stability. To the extent that soldiers who receive training resources leave the unit before going to the NTC or enter a unit after preparatory training, the effects of the training resources are lost.

PREDICTOR MODELS. There are a number of models possible for predicting unit performance, P, as explained below:

Simple Linear Prediction Model

$$P = A \text{ (TANK TRAINING MILEAGE)} + B \text{ (CARRIER MILEAGE)} + \\ D \text{ (TANK LIVE FIRE ROUNDS)} + E \text{ (BFV ROUNDS)} + \\ F \text{ (UCOFT HOURS)} + G \text{ (ARTBASS HOURS)} + \\ H \text{ (NTC TERRAIN SIMILARITY)} + I \text{ (AVERAGE STABILITY)} + \text{ERROR}$$

Interactive Model

The predictor variables specified above may be multiplied times each other in all of the possible combinations. Such a model would be complete, but exhaust the data set possibly available for testing the model.

Preferred Model

First, we form a cleaned-up criterion. The similarity of the NTC terrain to that available at home station for training is only of interest for its ability to account for variation in the performance criterion. In other words, if the training terrain rating were correlated with NTC performance and we were to ignore that relationship, prediction error would be needlessly high. Therefore, terrain similarity ratings will be used to predict performance; and the prediction errors, termed residuals, will be

retained as the cleaned performance criterion (an example of the successful application of this technique, with two extraneous predictors partialled out of the criterion, is found in Hiller, Fisher, and Kaess, 1969; Kerlinger and Pedhazur, 1973, elaborate on it also). The advantage of using the extraneous predictors to form residuals, over employing them as covariates, is to clarify the relationships between predictors of interest, such as tank and carrier training mileage, and the criterion.

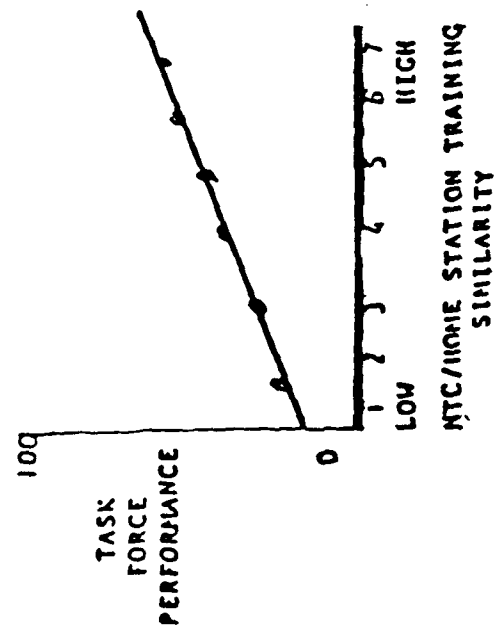
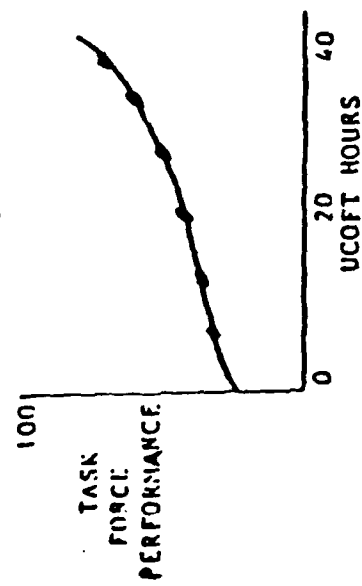
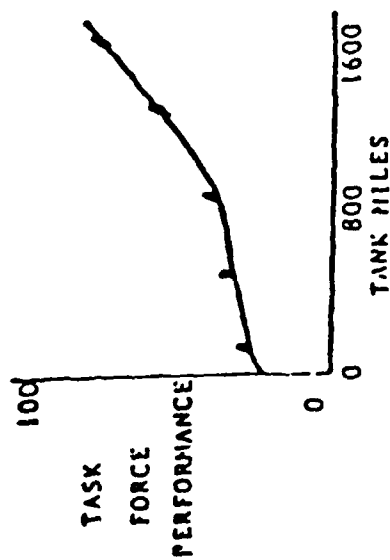
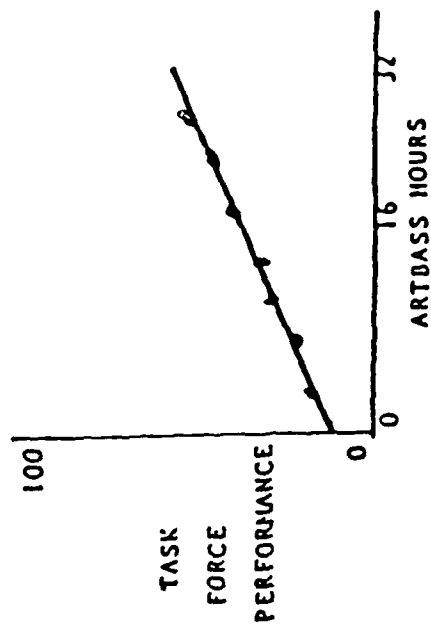
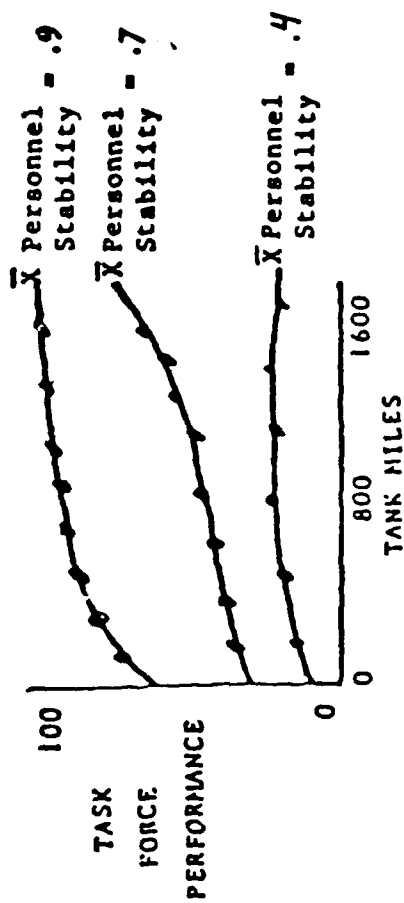
Second, consider that resources can contribute to training and performance acquisition/sustainment only if the unit's soldiers are present for training. Therefore, a measure reflecting a unit's stability should be multiplied times each resource measure to adjust its potential contribution. An index of stability is being formed on a monthly basis by matching the names of the soldiers who trained in units at the NTC with the names of the soldiers recorded in their unit's monthly personnel file for each of the six months preceding NTC. This will yield a measure describing the proportion of those present for training in each of the six months. These six proportions will be averaged for each unit and then used as a moderator variable. Hypothetical research findings using this analytic approach are graphed in Exhibit 1.

REFERENCES

- Hiller, J.H. (1987). Deriving Useful Lessons from Combat Simulations. Defense Management Journal, 2nd and 3rd Quarter, 28-33.
- Hiller, J.H., Fisher, G.A., and Kaess, W. (1969). Computer investigation of verbal characteristics of effective classroom lecturing. American Educational Research Journal, 661-675.
- Kerlinger, F.W. and Pedhazur, E.J. (1973). Multiple Regression in Behavioral Research. New York: Holt, Rinehart, and Winston.

EXHIBIT 1.

HYPOTHETICAL RESEARCH FINDINGS



Brigade Performance Measurement System

James T. Root
The BDM Corporation

For the soldiers of the 120th Infantry Regiment around the town of Mortain, France, on 6 August 1944, the day started early. German reconnaissance elements began bumping into the forward line just after midnight with the main attack arriving an hour and a half later. Their objective was to cut the supply lines for Patton's Third Army which had just been launched into France. Between the Germans and their objective was the 120th, so to ensure success, the Germans attacked with three tank divisions. Six days later, the Germans backed off, leaving the 120th still in control of Mortain and the Third Army still charging across France.

Why was the 120th successful? It's hard to say. At the time, the Germans were certainly not prepared to offer any helpful comments. On the American side, there was plenty of discussion, but it was only from the survivors, most of whom were too busy to sort out the how and why.

In more recent years, the National Training Center has incorporated brigade level exercises into their training program. Using an experienced Observer/Controller cadre and real time instrumentation, immediate training analysis and feedback of key events at Brigade level is now possible.

However, no mechanism currently exists to structure that information into a functional database available for technical and systemic military research. This paper outlines such a mechanism, one that has a continuing research application, as well as an immediate training value.

Project 2A was designed to assess battalion task forces, company/teams, and platoons by using a Critical Task Measurement System (CTMS) that explains the how and why of their respective performance.

The system is organized into three operational phases of planning, preparation, and execution that follow the flow of a battle. Each of these phases is further organized into segments that reflect the sequential flow of the phase. This framework is shown in Figure 1.

The mission critical tasks appropriate to each segment are then ordered and functionally linked as they relate to each other and tasks in adjacent segments. As an example, once a battalion commander is given a mission, he must conduct a mission analysis, understand his commander's intent, and understand the control measures before he can initiate his planning process. This linkage for the planning phase is illustrated in Figure 2.

Using the National Training Center as the system laboratory, CTMS has been developed for task force, company, and platoon for the following missions: Deliberate Attack Day, Deliberate Attack Night, Hasty Attack, Movement to Contact, and Defend. The next step is to extend this system to brigade level.

OPERATIONAL MODEL

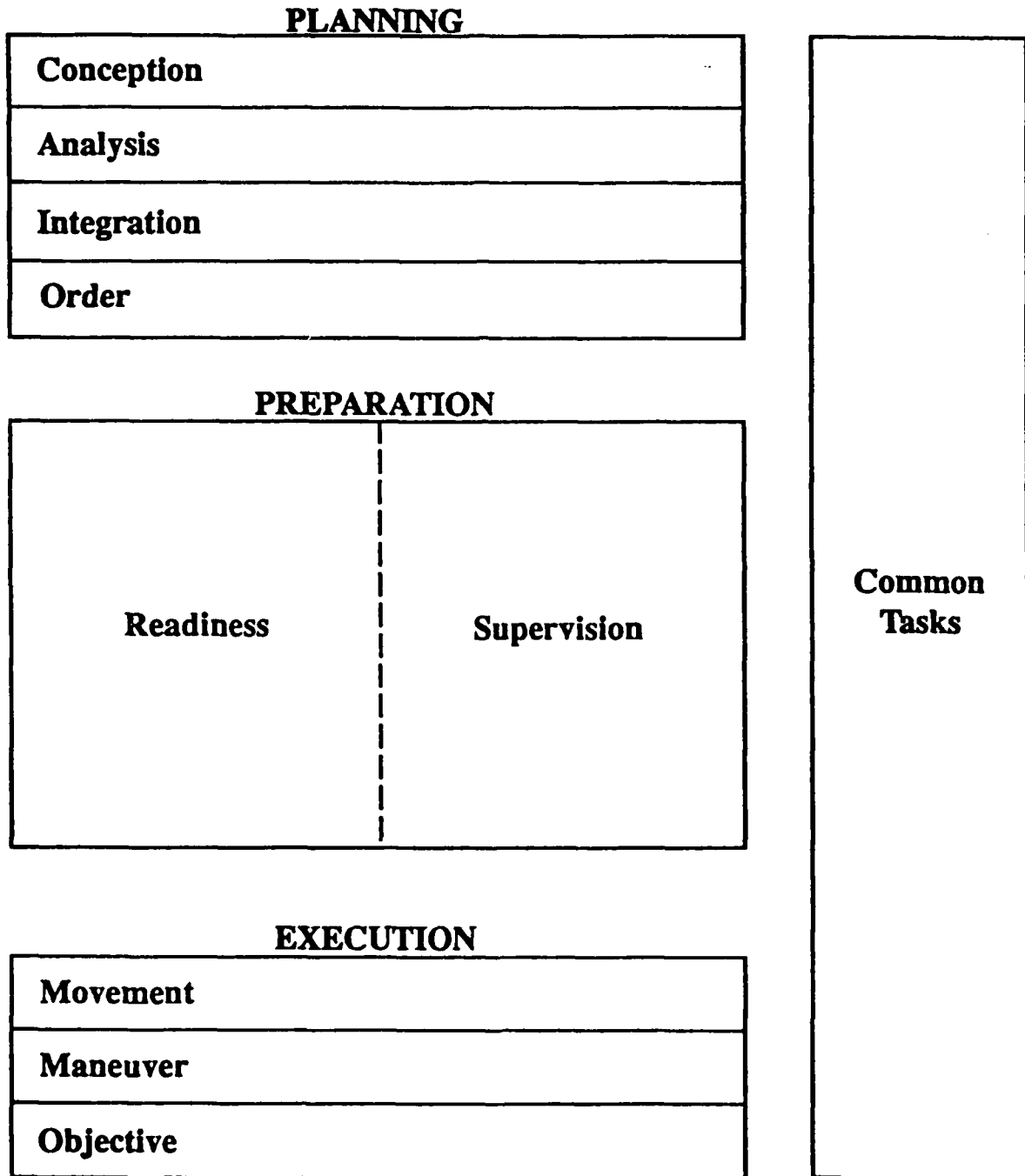
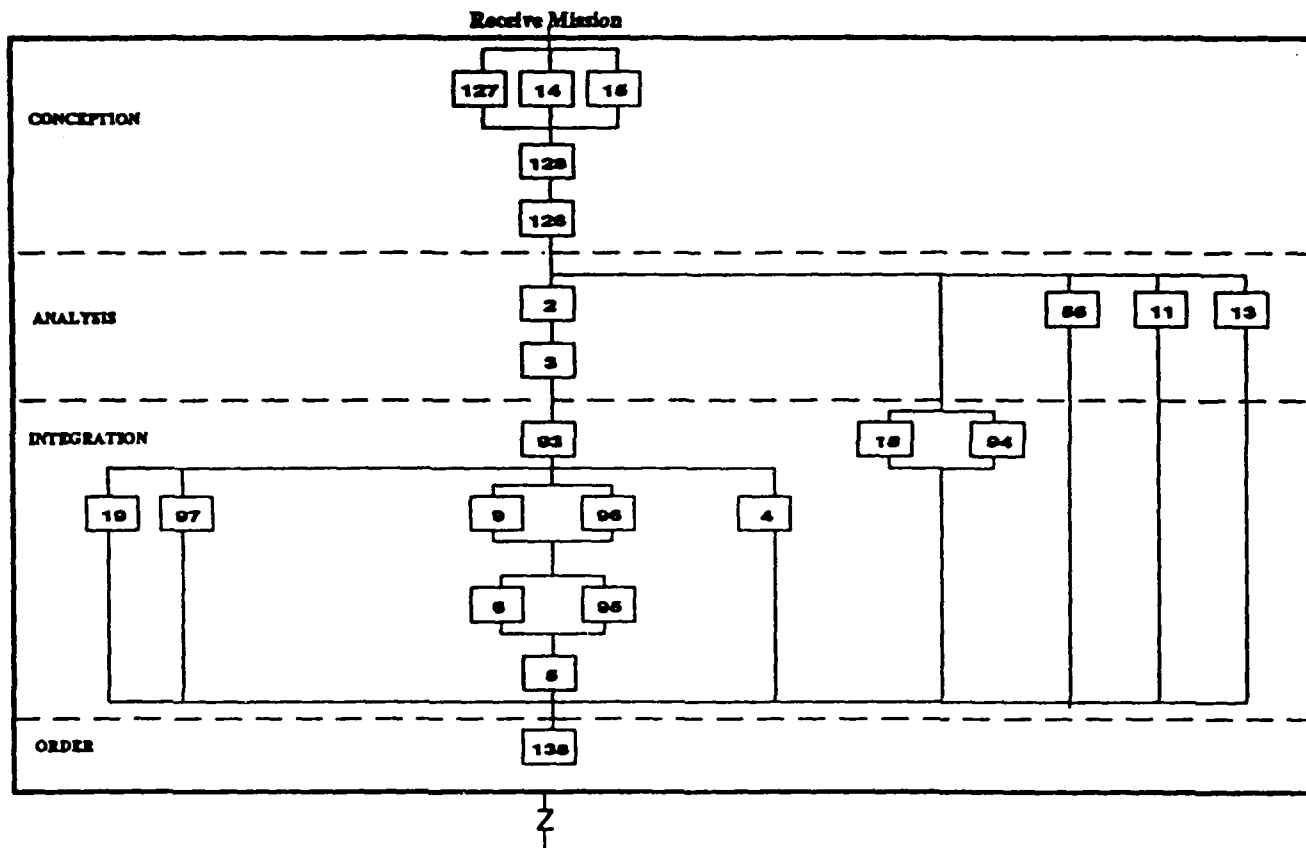


Figure 1.

FORMAT WITH TASKS

PLANNING SEGMENT SEQUENCING

PLANNING



MISSION TASKS

- | | |
|--|---|
| 2 Conduct Terrain Analysis. | 19 Plan Fire Control and Distribution Measures. |
| 3 Identify Enemy Strengths and Weaknesses. | 56 Plan Evacuation Procedures. |
| 4 Plan for Mutual Support. | 93 Coordinate Plans with Lateral Units. |
| 5 Plan Movement. | 94 Maintain Communications. |
| 6 Plan Actions on Contact. | 95 Designate a Support by Fire Element. |
| 9 Plan Reorganization. | 96 Designate Consolidation Procedures. |
| 11 Plan Air Defense Measures. | 97 Verify Supporting Fires. |
| 13 Plan for NBC operations. | 126 Issue Warning Order. |
| 14 Understand Commander's Intent. | 127 Conduct Mission Analysis. |
| 15 Understand Control Measures. | 128 Initiate Planning Process. |
| 18 Plan Redundant Communications. | 138 Issue FRAGO. |

Figure 2.

Before the model can be transferred to brigade level, however, it is first necessary to understand how a brigade operates. In this regard, the brigade has several unique characteristics not found at subordinate echelons. These characteristics can be understood through the operational concept of the 'Brigade Slice'.

This 'slice' represents those division and corps assets given to a brigade to meet mission requirements. Without going into detail regarding the various support and command relationships, it is sufficient to say that all these elements serve the needs of the brigade and its subordinate maneuver battalions. A normal slice contains the following:

- 1) Air Liaison Team (TACP)
- 2) Smoke platoon
- 3) Military Police platoon
- 4) Signal platoon
- 5) Air Defense company
- 6) Engineer company
- 7) Combat Electronic Warfare and Intelligence (CEWI) company
- 8) Attack Helicopter company
- 9) Artillery Direct Support battalion
- 10) Forward Support battalion, which includes a Medical company, a Supply company, and a Maintenance company.

These forces are organized into a Brigade Support Area (BSA) which also houses the Field Trains from the maneuver battalions. As the maneuver battalions are task organized for specific missions, elements of the brigade slice are sliced down to them. However, many of these elements remain under brigade control for the duration of operations. Figure 3 portrays the operational employment of brigade slice assets to a subordinate task force to meet mission requirements.

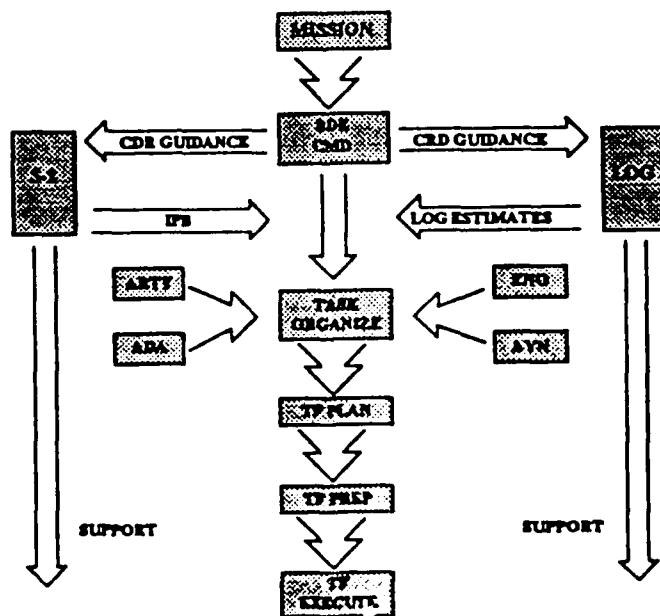


Figure 3

Brigade mission may not involve the entire brigade or they may involve different subordinate elements in different missions. For example, one task force, as an economy of force measure, may be in a defensive posture, while another task force is involved in a deliberate attack. This would result in a different mix and priority of support assets allocated to each respective task force. The remaining brigade support slice would continue its overall mission, adjusting as necessary for task force operational demands.

This semi-permanent support operation suggests a structural modification of the task force Critical Task Measurement System (CTMS). While the task force CTMS could follow a direct mission format from planning through execution, the brigade level format will necessitate an additional continuous level which feeds into the mission requirements. This structural modification is represented in the circular grid beside the CTMS format discussed earlier. (See Figure 4)

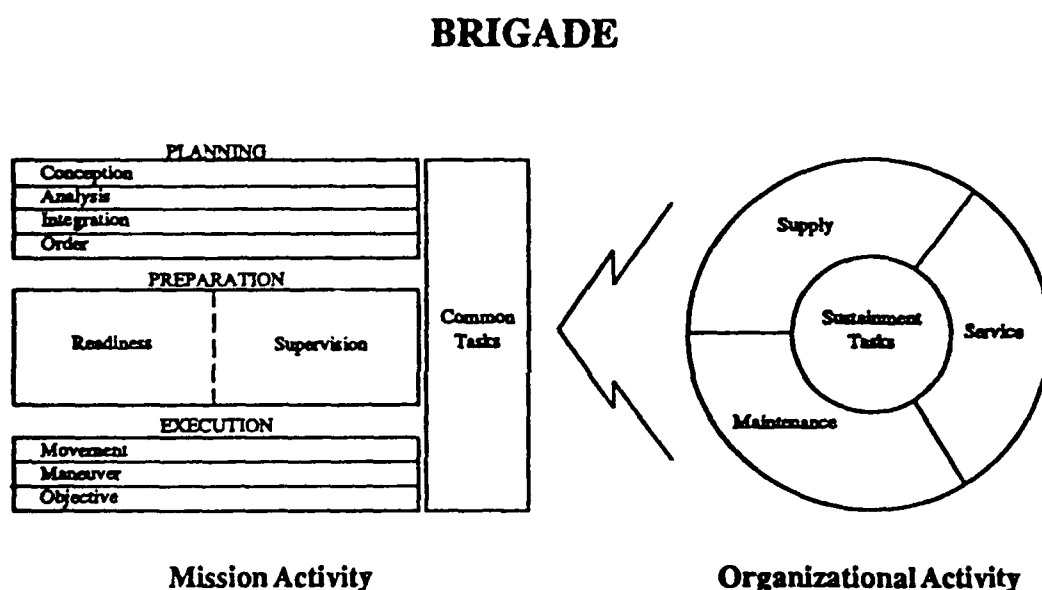


Figure 4.

It is clear that the mission specific event flow is coupled with sustainment operations that directly influence the conduct of the mission. It is, therefore, envisaged that the linkage from brigade to echelons above and below will be through the Battlefield Operating Systems. In this way, the processing of information through the task linkage will develop a focused identification of systemic issues and their impact upon the respective echelons. Figure 5 illustrates this linkage.

Division Support Linkage

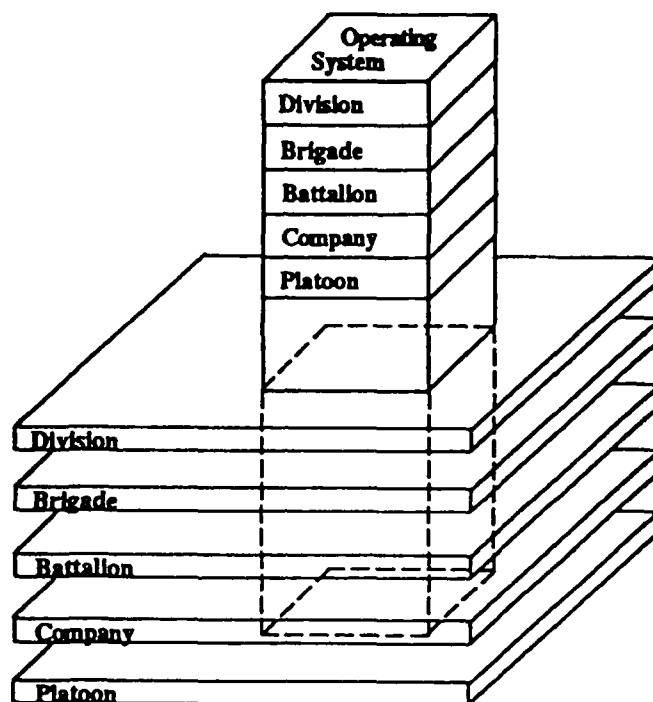


Figure 5

The brigade represents the critical link between the echelons that must actually fight the battle and the resource echelons that support the effort. The echelons below brigade can be assessed through the 2A methodology. The echelons above, specifically divisions, are being assessed through the Battle Command Training Program. Because brigade is the linkage between the two, it requires a mechanism that reflects that connection. By utilizing the 2A methodology and focusing on the battlefield operating systems, that assessment mechanism can be developed.

Assessing Light Forces at the Joint Readiness Training Center

Judith J. Nichols
The BDM Corporation
and
LTC Howard W. Crawford
Operations Group, Joint Readiness Training Center

The Army currently has four Combat Training Centers (CTCs) in various stages of operation. The oldest and most established CTC is the National Training Center (NTC) at Fort Irwin, CA. The Battle Command Training Program (BCTP) at Fort Leavenworth is in its first year of operation. The Combat Maneuver Training Complex (CMTC) at Hohenfels, FRG is in start-up mode, and the Joint Readiness Training Center (JRTC) has recently completed its first full year of operation.

Although each CTC has a different training focus, all CTCs share the same train-evaluate-train philosophy to include:

- (1) Highly realistic training scenarios;
- (2) Dedicated opposing force (OPFOR);
- (3) Professional Observer/Controllers (O/Cs); and
- (4) Systematic evaluation and feedback in the form of After Action Reviews (AARs) and a comprehensive Take Home Package (THP);

This paper focuses on the "evaluate" component of the train-evaluate-train approach to CTC training and addresses the development of unit performance assessment mechanisms (i.e., AARs and the THP) during the Joint Readiness Training Center's start-up and first year of operation.

Assessing Light Units at JRTC

The specialized nature of the units training at JRTC required an innovative approach to unit performance evaluation. The JRTC unit assessment mechanism had to ensure systematic evaluation of light infantry units, air assault units, rangers, air borne units, and special operations forces. Additionally, units rotating through JRTC were slated to conduct both low-intensity and mid-intensity combat scenarios.

The variety of unit types and the rotational transition from low-to mid-intensity conflict at JRTC presented a number of difficulties to systematic data collection and evaluation. In contrast to the NTC where training units and training missions are more or less generic, the JRTC

was faced with the problem of evaluating units that do not have a generic composition and do not necessarily execute the same missions.

In order to best serve the training needs of these diverse non-mechanized units at JRTC, a series of training and evaluation outlines was developed. These would become the foundation for training and evaluation at JRTC. They would first serve as the primary checklist for observer controllers in the field and secondly as a uniform standard for training OCs. Both were vital to a new organization trying to train trainers and build a uniform evaluation program concurrently.

JRTC Data Collection

Unlike the NTC where much of the data used to support unit feedback is captured through the NTC Instrumentation System, the JRTC -- which is not instrumented -- had to rely heavily on manual data collection. Figure 1 shows the various data sources and source-specific information routinely collected at JRTC.

<u>DATA SOURCE</u>	<u>INFORMATION COLLECTED</u>
OBSERVER CONTROLLERS	<ol style="list-style-type: none">1. T&EO checklist data and field notes2. Unit mission summaries3. Training observations4. Firing data5. BDA (personnel/equipment)
EXERCISE MANAGEMENT & CONTROL CENTER (EMCC)	<ol style="list-style-type: none">1. Operational Logs (Intel, ADA, Maneuver companies, Aviation, Engineer, CSS)2. Joint Logs (Air Opns)3. Fire marker control (BDA)
I-MILES TEAM	<ol style="list-style-type: none">1. Engagement data (AT weapons/ combat vehicles)2. BDA
VIDEO TEAM	<ol style="list-style-type: none">1. Video tapes (TF AARs, System AARs, selected CO/PLT AARs)
OPFOR	<ol style="list-style-type: none">1. BDA from each battle2. Sketches on each battle

Figure 1. JRTC Data Sources

The primary source of all unit performance data at JRTC is the Observer/Controller (O/C) Team. O/Cs collect and/or generate approximately eighty-five percent of all unit performance data. Once potential data sources had been identified, the JRTC developed a plan which provided systematic procedures and mechanisms to ensure data reliability.

Training and Evaluation Outlines (T&EOs). The primary mechanism for collection of JRTC unit performance data was directed by the Training and Doctrine Command (TRADOC) to be a series of Training and Evaluation Outlines which was doctrinally sound and in line with Army-wide Mission Training Plans (MTPs) currently under development at the schools. Working directly with the Center for Army Lessons Learned (CALL), the JRTC initiated a series of workshops with Army schools, CATA, and various proponent agencies to develop the Training and Evaluation Outlines to be used at JRTC. The goal of these workshops was to produce data collection instruments in the form of O/C checklists that would support light infantry performance evaluation. These workshops ultimately resulted in the identification of:

- (1) A mission essential task list (METL) for each of the 7 battlefield operating systems (BOS); and
- (2) A T&EO structure (format) based on Army doctrine (e.g., FM 25-100, TRADOC Reg. 310-2) and in line with USAIS MTP requirements.

The initial METL list served as the start point for T&EO production. Each mission and operating system was reviewed and a priority -- from most likely to occur at JRTC to least likely -- was established along echelons for maneuver missions at battalion, company, platoon, and squad levels. Once these priorities were fixed, production began on writing T&EOs. While many of the T&EOs were written at JRTC, much help was received from the various schools or, when possible, existing MTPs were used as the base for the T&EO. In all cases, these initial T&EOs, approximately 160, were approved by the appropriate school house in time for the first rotation. Figure 2 illustrates both the content and format of a current JRTC T&EO.

There is a separate set of T&EO checklists for each battle field operating system (i.e., maneuver, intelligence, fire support, combat service support, air defense artillery, mobility/ countermobility/survivability, combat support). Additionally, there are T&EOs for the aviation and close air support (CAS) components of the task force.

TASK TRAINING STATUS:

T P U

ITERATION 1 2 3 4 5

CONDITIONS: The battalion has been given a Bde/Reg OPORD to move from their present location to another area, occupy it, and prepare for combat operations or, the battalion has just completed combat operations and must move to a relatively secure area.

TASK STANDARDS:

1. The unit maintains security throughout the operation so that the main body is not surprised by the enemy.
2. The unit does not sustain any casualties by friendly fire.
3. The unit establishes and maintains command & control throughout the operation.
4. The battalion completes all preparations as specified in the order.

SUBTASKS AND STANDARDS:

GO NO GO NA/NO

1.	In offense, was site selected well forward to avoid early displacement?	_____	_____	_____
2.	In defense, was site selected well rear to avoid displacement and minimize arty/EW?	_____	_____	_____
3.	The selected assembly area had all the necessary characteristics.	_____	_____	_____
3a.	Cover and concealment?	_____	_____	_____
3b.	Room for dispersion?	_____	_____	_____
3c.	Security from air and ground attack?	_____	_____	_____
3d.	Covered routes in and out?	_____	_____	_____
3e.	Situated on defensible terrain?	_____	_____	_____
3f.	Out of enemy direct fire and light indirect fire range?	_____	_____	_____
4.	The quartering officer accomplished all the necessary tasks.	_____	_____	_____
4a.	Select exact CP site?	_____	_____	_____
4b.	Designate platoon position?	_____	_____	_____
4c.	Post guides to lead incoming elements?	_____	_____	_____

Figure 2. JRTC T&EO Task 18: Occupy an Assembly Area (Bn)

Data Analysis

Checklist data collected by O/Cs at all echelons is compiled and analyzed by each operating system and slice element/component O/C team chief and briefed to the senior O/C in preparation for the unit After Action Review. One of the benefits accrued from T&EOs is the ability to annotate both leader and critical tasks. This allows OCs to quickly focus on the most important tasks and subtasks in the T&EO. Battle damage assessments (BDA) and firing data also are compiled and provided to the senior O/C and his AAR staff for use in the evaluation. There are a number of other data sources included in the data analysis process, specifically I-MILES data and appropriate extractions from the EMCC logs. All data are turned in to EMC/Plans at the close of each mission phase.

JRTC Unit Evaluation and Feedback

The After Action Review is the primary mechanism for JRTC unit evaluation and feedback. AARs are conducted at the close of each mission phase. A mission phase is an extended period of time -- approximately 72 hours -- during which various related tasks (e.g., occupy an assembly area, move tactically, assault the objective, consolidate and reorganize) are executed to accomplish a broader mission (conduct a deliberate attack). AARs are conducted at the battalion, company/team, and platoon echelons at the close of each mission phase. System AARs are also conducted for CSS, Fire Support, Aviation, and Intel at intervals throughout the rotation.

Units are also provided with a complete record of their performance at JRTC in the form of a Take Home Package which includes operating system trends, mission summaries, task performance evaluations, identified unit strengths and areas in need of improvement, and follow-on training recommendations. THPs also include video tapes of all battalion level, slice element, and selected company/team and platoon AARs.

Summary

The development and implementation of unit performance evaluation mechanisms at the JRTC reflect the CTC train-evaluate-train philosophy. Light forces rotating through JRTC undergo a series of training exercises that include both low- and mid- intensity combat missions which are evaluated at the close of each mission phase.

In order to meet the needs of each type of unit training at JRTC, a series of training and evaluation outlines was developed which provided the basis for the compilation of O/C checklists which address all special unit and mission requirements. There is a separate set of T&EO checklists for each battlefield operating system (i.e., maneuver, intelligence, fire support, combat service support, air defense artillery,

mobility/countermobility/survivability, combat support). Additionally, there are T&EOs for the aviation and close air support (CAS) components of the task force.

AARs are conducted at the battalion, company/team, and platoon echelons at the close of each mission phase. Formal AARs are also conducted for the certain systems (CSS, FS, Aviation, Intel) at intervals throughout the rotation. Units are also provided with a complete record of their performance at JRTC in the form of a Take Home Package which includes operating system trends, mission summaries, task performance evaluations, identified unit strengths and areas in need of improvement, and follow-on training recommendations.

MEASUREMENT ISSUES AT THE
BATTLE COMMAND TRAINING PROGRAM (BCTP)

William A. Ross and Karol Girdler
The BDM Corporation

This paper explores the nature of the measurement issues characterizing the development of the Battle Command Training Program (BCTP) during its initial 1 and 1/2 years. BCTP is the U.S. Army's newest operational Combat Training Center (CTC) designed for collective training at the Division and Corps level with an emphasis on decision making skills among senior commanders and staffs. The measurement issues, like BCTP, have evolved over the initial development period. The issues are, however, still far from final resolution due to both the complexity of development typical to the CTCs, and due to the unique nature of doctrine and training at Division and Corps. Due to the echelons training at BCTP, the measurement lessons learned during previous CTC development only partially apply. The issues which will be discussed concern 1) what to measure, 2) how to measure, 3) how to consolidate data from various measurement sources and produce performance feedback, and 4) constraints to programmatic approaches for the refinement of measurement issues. The issues are presented within a framework representing the flow of information through the system. See Figure 1.

Figure 1.
Measurement Issues

<u>INFORMATION FLOW</u>	<u>WHAT TO MEASURE</u>	<u>HOW TO MEASURE</u>	<u>DATA CONSOLIDATION AND PRESENTATION</u>	<u>PROGRAMMATIC REFINEMENT</u>
PURPOSE OF DATA COLLECTION	X			X
DATA COLLECTION STRATEGY	X	X	X	X
DATA COLLECTION METHODS- IMPLEMENTATION		X		X
ANALYSIS STRATEGY	X		X	X
ANALYSIS METHODS- IMPLEMENTATION			X	X
PRESENTATION METHODS			X	X

Both the complexity of the issues at Division and Corps, and the constraints (time and manpower) on BCTP led to the initiation of an effort by BDM to begin an integrated resolution of the measurement issues. This effort will be referred to as it pertains to each issue below.

The resolution of measurement issues at BCTP has been affected by the multiplicity of data needs and of change which characterize both the internal and external environments. These effects are especially evident in the area of what to measure. The nature of measurement at Division and Corps is heavily tied to command and control which is illusive at best, and difficult to tie to outcomes. In addition, the development of BCTP is tied to a number of innovations in computer simulation technology which can change the nature of the events being measured from exercise to exercise. Philosophical differences exist within the community about the relative measurement emphasis on outcomes of battle versus emphasis on the staff processes during the battle flow and the relationship of the two for diagnostic purposes. In addition, doctrine, a basis for what to measure, has been in flux at Division and Corps, with many new publications being written and reviewed during the past 1 and 1/2 years. Among these were original work on both Division and Corps Army Mission Training Plans (AMTP), and rewriting of basic Field Manuals (FMs). Central to a resolution of what to measure is the assumption that what to train is well understood. Given the changes in doctrine during the past 1 and 1/2 years, what to train and what to measure have been difficult to determine doctrinally to the level needed for the design of measurement tools.

In addition to deciding what to measure about unit performance, BCTP has been faced with the need to construct an elaborate system to gather data on the flow of the battle for exercise management purposes. Due to the nature of the Joint Exercise Simulation System (JESS) model used to drive the exercises, a great deal of data was needed by the exercise director. The nature of the need for exercise management data is that of a view of the unfolding battle and specific activities so that the exercise director can determine that the exercise is indeed serving the training objectives, i.e., stimulating the training audience and not distracting them with unrealistic portrayals. Additionally, some manually generated (scripted) events must be integrated into the flow of the computer driven exercise. BCTP, like all CTCs, has a training mission as well as responsibility for providing the Army with data to drive Lessons Learned. To meet this mission for systemic feedback, masses of data had been collected at each exercise and archived by the Army's Center for Lessons Learned (CALL). What to measure was undefined. Lack of articulated data collection and analysis strategies further meant that the ultimate purpose of data collection was not being met. Conclusions were impossible to derive from the masses of raw data.

Temporary solutions for what to measure to meet these diverse needs have included a reliance on command emphasis,

attempts to measure "everything," and an emphasis on using resources to gather data to insure that the exercise reasonably stimulated the training objectives and did not unreasonably distract the training audience. As the Division AMTP evolves, the external environment is forcing a new emphasis on a fine grained analysis of unit performance, and as BCTP matures, a more defensible measurement and feedback plan is needed to maintain credibility. It is at the point in BCTP program evolution that the BDM effort, the Warfighter Feedback System (WFS), is intervening.

The approach to what to measure taken by the BDM project, the Warfighter Feedback System (WFS), was to base the data collection and feedback more closely on the current and most stable doctrine, and to synthesize the emerging elements of a good feedback system which BCTP had already developed. Some principles and needs of BCTP must define the purposes of data collection and provide a framework for resolution of the first issue. These purposes can be summarized as follows:

First, BCTP defines the uniqueness of its program in terms of diagnostic feedback to the unit. The primary purpose of data collection is to provide the participating unit with feedback on its effective or ineffective performance. Second, BCTP has a commitment to package the feedback in terms of the Battlefield Operating Systems (BOSs). Third, the data must capture something about the efficiency of the unit. Fourth, areas of the simulation which can cause negative learning or distraction must be monitored closely by assessment of particular data elements. Fifth, the requirement must be met to provide data useful as systemic feedback.

The issue of what to measure in order to meet these various needs is not resolved. The use of organizing frameworks to consolidate the BCTP data collection purposes into useful strategies is underway. One effort is the organization of the critical collective tasks for each BOS according to a framework which makes sense for that BOS. To date, the collection framework for the Intelligence BOS has been completed. It was found that doctrine in this area was complete and stable enough to define what to measure. This effort built on BCTP's commitment to provide feedback within the BOS framework. Data collection following this organization has been tested. The second effort concerns the construction of an issue based data collection strategy. This strategy organizes the elements of the macro-tasks of Divisions (e.g., counterfire). The elements of each issue are linked to existing data sources to provide a means of aggregating the masses of data into useful feedback. The data collection and analysis strategies can then become synchronized. Although the issue based method had been used by BCTP in its initial exercise, the requirements for data collection had never been articulated in a manner which caused a routine understanding of what to collect.

Given some organization about what to collect, there is

still a significant gap to derive useable collection methods. Several sources of data exist at BCTP. There is no helpful instrumentation. The JESS data were limited due to the lack of a computerized post-processor. A number of people manning work stations attempt to capture information about battle events as they occur. They provide narrative summaries, as well as summaries of battlefield effectiveness such as artillery success. The most important data collection resource is the cadre of Observer/Controllers (OCs) which is required to cover seven or eight geographically dispersed sites 24 hours per day. The best use of OC time was hampered by the lack of a data collection strategy and measurement methods to carry out the strategy. In addition, methodology to synthesize the collection efforts of these two primary sources, JESS through workstation controllers and OCs, was missing.

The BDM effort is attempting to provide the understanding of and to support the development of a strategy which synthesizes the qualitative and quantitative data available from these sources. BDM's project at BCTP is based on an INTEGRATED SCIENTIFIC PARADIGM for data collection. It includes structured approaches for gathering both qualitative and quantitative data. This integrated paradigm offers several advantages. First, it provides a strategy for deriving a fuller and more complete picture of unit performance that is possible by relying on a single methodology. Second, a more meaningful and credible perspective on unit performance can be built by combining quantitative indicators (such as counts and percentages) with qualitative accounts (supported by observable evidence). Third, using a structured qualitative framework that makes explicit the differences between reasoned judgments and supporting evidence takes narrative accounts by OCs out of the realm of "subjective" data and into a scientifically defensible methodology for describing human behavior.

BCTP is the first CTC to attempt using an integrated scientific paradigm for data collection. Currently, feedback at the National Training Center (NTC) and the Joint Readiness Training Center (JRTC) relies on quantitative indicators (for example, from the NTC instrumentation system or BDA estimates provided by OCs) and on written narrative accounts by OCs incorporated into the Take Home Package. However, these narrative accounts are not gathered within a systematic qualitative framework which leaves them open to charges of subjectivity. Furthermore, quantitative indicators are typically treated separately from narrative accounts resulting in a lack of integration (and occasionally leading to conflicting conclusions).

BCTP's leadership in using state-of-the-art scientific methodology for collecting data on unit performance makes the tryout of the OC Guidebook particularly important. The results of the tryout will not only have implications for BCTP's data collection system but may provide a model for other CTCs (especially the Combat Maneuver Training Center currently coming

on line).

To support the development of this perspective, BDM has developed data collection forms for the OCs which combine data types. Additionally, we are working with the data flow during exercises to provide recommendations for a smooth flow of data from collection to analysis to feedback. To see the usefulness of the data collection strategy and specific methods, the outcome of analysis and feedback must be evident to the collectors. This added context should improve the quality of data collected.

BCTP has recently made gains in its ability to consolidate and present the data to the training audience. These gains are based on the evolving data collection issues as a collection strategy and modeling of appropriate analysis by key BCTP personnel. Key elements in their use of data for feedback were the improvements in their After Action Review (AAR) preparation procedures and their more rigorous application of their own model of feedback, i.e. linking observations with outcomes of unit effectiveness. Observations of unit efficiency which were difficult to link to outcomes were also given a more clearly articulated place in the feedback process.

There are three or four AARs during the five-day BCTP Warfighter exercises. The AAR preparation period begins with the OCs providing a backbrief to the BCTP Commander about the data they have collected. This backbrief has become the data reduction and analysis work period where presentation priorities are also discussed. Recently, in one Warfighter Exercise, the issue based collection and analysis strategy was used as the meeting framework. At the same time, workstation controller data were introduced directly into the meeting for examination with the OC data. A model for presentation helped further analyze the data. This model was that observed performance must be clearly accompanied by related performance outcome data. This type of data would be presented in the AAR. Observations of inefficient staff performance not clearly linked to ineffectiveness were noted and assigned as material for the "informal AARs" held at the Command Posts (CPs) by the OCs. The use of these methods in the meeting were effectively modeled by the Commander of BCTP.

The use of the issue based strategy, overlaying the BOS framework, defining observations as relevant to efficiency or effectiveness, and the introduction of data from another source provided clearer cognitive maps for the OCs, the primary data collectors. The purposes and uses of data collection were clearer. The flow of the battle which had unfolded was also clearer.

To complete the feedback system, the AAR format was revised in that exercise to more clearly highlight the observations and their relation to outcomes. Extraneous portions of the AAR which had previously diffused the focus on the learning points were omitted. Given the imperfect nature of field data, the unit was requested to thoroughly discuss observations from its viewpoint.

Use of these emerging organizing strategies in one exercise for a successful AAR does not mean that the measurement issues are resolved. Rather, the approach must be held consistently long enough to develop the strategies and attendant methodologies. The macro-task issues have not been defined so that data collectors can have consistent use across settings. Measurement methods relative to the various issues are under development.

The complexity of action and measurement at Division and Corps requires a network of organizing structures across the information flow framework. Success at BCTP can only be obtained by investigating the multiple organizing structures thoroughly, consistently maintaining a structure until implementation methods can be developed, tested, and revised, beginning to institutionalize with training, and realizing successful feedback experiences. For these actions to happen, BCTP leadership must contend with a number of constraints and distractions. The environment could cause the development of measurement feedback at BCTP to lose focus and flounder. BCTP is currently prone to multiple internal innovations outside of an organizing context or effective evaluation program. It is easy to abandon innovations and developments and to heap together too many changes causing a loss of cause and effect diagnosis. No integration mechanism exists for controlling innovation. Resolution of these issues is immensely complicated, but possible given a focused approach which links concepts to methods, implementation, evaluation and revision.

USING SIMULATED NETWORKING TECHNOLOGY AS A COMBAT TRAINING CENTER SURROGATE

William J. Doherty & Nancy K. Atwood
The BDM Corporation

Introduction

The need for Army readiness has never been more acute than at the present time, yet in this era of deficit reduction the opportunity for resource-intensive training is being greatly curtailed. This paper examines the role of simulated network technology (SIMNET) as a possible surrogate for Combat Training Center (CTC) advanced collective training. The paper is organized into four sections: a description of the CTC program and its place in Army Training, a description of SIMNET, a discussion of potential training strategies for SIMNET, and finally, a discussion of the potential role for SIMNET relative to the CTC training.

The Combat Training Center (CTC) Program

Inherent in the training cycle and use of the AMTP is a train-evaluate-train model which allows for ongoing diagnosis and remediation of training needs. A major component of forces training is the CTC (Combat Training Center) program. This program presently consists of four separate components:

- (1) The National Training Center (NTC);
- (2) The Joint Readiness Training Center (JRTC);
- (3) The Combat Maneuver Training Center (CMTC); and
- (4) The Battle Command Training Program (BCTP).

The purpose of this program and its four elements is to provide realistic combat-like training that units cannot achieve at their own home station. The NTC, which has been operable for more than seven years, focuses primarily on mid- to high-intensity combat training for heavy mechanized units. The JRTC, which began training rotations in the fall of 1987, provides low- to mid-intensity joint combat training for light forces (light infantry, airborne, air assault, and special operations forces). The CMTC, currently under development, will provide a comparable experience to that at the NTC but for units stationed in Europe. The BCTP, which ran its first warfighting exercise in January 1988, is designed to use computer simulations to provide command and battle staff training at the Division (and eventually Corps) level, relying on computer simulations.

The importance of the CTC program for forces training cannot be overstated. These programs provide the most realistic opportunity for the unit to exercise in an environment similar to that of its wartime

mission. As a result, not only does the unit accrue valuable experience in performing its missions and tasks in a stressful environment but it is also possible to assess the combat effectiveness of the units in a more objective and meaningful way than is available at home station.

The recent draft of AR 350-50 establishes policies, objectives, and responsibilities for the Combat Training Center program. The five objectives of the CTC program are shown in the following Table 1.

TABLE 1. CTC PROGRAM OBJECTIVES

-
1. INCREASE UNIT READINESS FOR DEPLOYMENT AND WARFIGHTING
 2. TRAIN BOLD INNOVATIVE LEADERS THROUGH STRESSFUL EXERCISES
 3. EMBED DOCTRINE THROUGHOUT THE TOTAL ARMY
 4. PROVIDE FEEDBACK TO ARMY AND JOINT/COMBINED PARTICIPANTS
 5. PROVIDE A DATA SOURCE FOR LESSONS LEARNED AS AN AID TO IMPROVE DOCTRINE, TRAINING, ORGANIZATION, MATERIEL, AND LEADERSHIP
-

The concept of operations for the CTCs is based on three pillars of advanced collective training:

- (1) A dedicated, doctrinally proficient Operations Group;
- (2) A dedicated, realistic OPFOR; and
- (3) A training facility which closely replicates combat conditions, and uses a system of instrumentation designed to unobtrusively collect and record battle events for replay and analysis.

AR 350-50 (Draft) explicitly identifies the dual role of performance assessments at the CTCs. The regulation directs that feedback from performance assessments is provided to the training unit and the unit's chain of command in the form of After Action Reviews (AARs) and written Take Home Packages (THPs). Furthermore, it calls for a trend

analysis of data from the CTCs which does not identify specific units. This analysis is intended to yield four primary outcomes:

- (1) Input to the Concept Based Requirements System;
- (2) Feedback to the total Army;
- (3) Feedback to the Joint Center for Lessons Learned; and
- (4) Support feedback to units training at CTCs.

The centrality of unit performance measurement at the CTCs is readily apparent from the Army's overall approach to training (FM 25-100) as well as the specific objectives of the CTC program itself (AR 350-50). However, design of successful measurement, analysis, and feedback strategies depends on a thorough understanding of the operations of each of the CTCs. Further, if SIMNET is to act as a surrogate for a CTC, it is imperative that a Unit Performance Measurement System comparable to a CTC, be employed in its training application.

There are several critical features of the CTC environment that contribute to a realistic combat simulation:

- (1) A free-play training environment;
- (2) A dedicated Opposing Force which uses appropriate threat tactics and equipment;
- (3) A battlefield environment which includes electronic warfare and close air support;
- (4) Extensive use of tactical engagement simulation, including MILES (Multiple Integrated Laser Engagement Simulation); and
- (5) A complex instrumentation system for real-time data collection of mission events.

SIMNET's ability to provide each of these elements is addressed in the last section of this paper.

Simulated Networking Technology (SIMNET)

The SIMNET program is a Defense Advanced Research Program Agency (DARPA) sponsored research effort designed to demonstrate the feasibility of developing a low cost simulator which when coupled with networking technology can be used to simulate collective performance in combat conditions. This research has resulted in the development of simulators for both ground vehicles (M-1 tanks, and Bradley Fighting Vehicles) as well as fixed wing and rotary aircraft. Currently, groups of these simulators have been installed at Fort Knox, Fort Benning, and in Europe. The installation at Fort Knox will eventually include a Battalion size

complement of simulators. A separate facility at Fort Knox includes other SIMNET simulators designed to be used for developmental purposes rather than training purposes. It is anticipated that this developmental effort will greatly reduce costs associated with new weapon systems or their modifications.

The simulators at a single locale are networked together using a Local Area Network (LAN) to share data packets allowing them to "see" each other on the same terrain data base. Eventually through the use of "long Haul" networking technology it will be possible for units at Fort Knox to share the same battlefield as units at other locations such as Fort Benning or in Europe.

The Army is currently examining the role of SIMNET and its production version, the Close Combat Tactical Trainer (CCTT), in terms of its place in the entire Army training system. The Army Research Institute through its Fort Knox Field Unit is supporting this effort by a research contract that will develop both a unit performance measurement system concept as well as an overall training strategy for SIMNET. Progress on that effort was the topic of an earlier session of this conference.

SIMNET Training Strategy

In understanding whether SIMNET can be used as a surrogate for the CTCs, it is necessary to place SIMNET in the Army training system environment. As a training device, SIMNET can be viewed as offering the seven advantages that any training device offers (Beecher, 1987): cost savings on fuel, mileage, and parts; impeding the forgetting period; permitting training on dangerous procedures at no risk; providing year-round training; avoiding environmental constraints; providing diagnostic feedback; and predicting performance on actual equipment. While the above provides a short list of benefits to be derived it does not provide a description of the role of SIMNET as a trainer.

The current TRADOC System's Manager (TSM), COL Mengel sees the training concept for the production version of SIMNET has being based on five tenets:

- (1) It is a unit trainer dedicated to collective training tasks and missions.
- (2) The chain of command is in charge of training.
- (3) Training is always in a combined arms environment. All of the Battlefield Operating Systems (BOS) as fully represented.
- (4) Every participant is a trainee.
- (5) Units train to their Mission Essential Task List (METL).

COL Mengel goes on to indicate that SIMNET fits into a combat unit's overall training program in two modes: as a precursor to tactical field training, and as a trainer of tactical operations that cannot be trained in the field because of their risk or expense.

In this view of SIMNET's training role, it is evident that SIMNET is seen as part of the Army's collective training system, and that it should provide many of the same features as the CTCs provide in this system.

SIMNET as a Surrogate CTC

As indicated above, there are five elements that combine the near real combat environment of a CTC: free play, dedicated OPFOR, electronic warfare and Close Air Support, use of tactical engagement simulation, and a sophisticated instrumentation system. Examining SIMNET with respect to these five dimensions, allows for an assessment of how closely it can reproduce a CTC environment. SIMNET as it is currently used by the troops at Fort Knox generally is in a free play environment. That is the performance of the unit and crews is responsive to the actions as they happen in the simulation. They are not preprogrammed or scripted. SIMNET is also capable of supporting electronic warfare and close air support. The simulation, through the use of audio and visible stimulation, provides a comparable experience to that provided by tactical engagement simulation. Vehicle losses are visibly noted, and when one's own vehicle is hit or killed this event is also simulated. Finally, like a CTC instrumentation system, vehicle status data is captured throughout the exercise so that replay or analysis can be applied at the completion of the exercise. In fact, the data captured by SIMNET in this regard is of a finer level of detail than that currently captured at the CTCs.

The only area where SIMNET may not fully match the CTC environment is the dedicated OPFOR. In SIMNET, it is possible to play other individuals in simulators or to play against an semi-automated OPFOR. The first case is similar to that at a CTC, however, there is no standing OPFOR cadre of players at each SIMNET facility. Thus, free play against an OPFOR is currently being conducted against other training troops likely using US tactics. The semi-automated OPFOR simply reduces the manpower requirements for this play. The lack of an OPFOR is a serious limitation when considering SIMNET as a potential surrogate of CTC training and performance. As COL(P) Wesley Clark, Commander of the BCTP program and former Chief of the Operations Group at NTC maintains, the OPFOR provides the standard against which US troops perform at a CTC. Therefore, if the Army considers it desirable to use SIMNET as a surrogate CTC, it will be necessary to address this important issue. Perhaps, as with the BCTP program, the use of a Soviet replicate (either semi-automated or fully automated) OPFOR would be possible.

While the above discussion has looked at the elements that comprise the CTC environment, it is important also to recognize the three pillars on which the CTCs are based: an Operations group, a dedicated

OPFOR, and a sophisticated instrumentation system. The latter two components have been addressed above. The Operations Group at a CTC provides the training that a unit receives at a CTC. Specifically, the Operations Group plans the scenarios, manages the exercises, evaluates the training, and provides the feedback. These functions must be replicated for SIMNET if it is to approximate the kind of training that units receive at the CTCs. In particular it is necessary to create specific scenarios that stress specific collective tasks at target echelons in order to insure that the units are training on appropriate AMTP tasks. Further, these scenarios need to be accompanied by an exercise management control system that insures that the training exercise is being conducted as planned. Finally, there needs to be provision for individuals to evaluate the training and perform the feedback function. At the CTCs, this provision is handled through dedicated Observer/Controllers and an associated feedback support cell. Presently, there are no comparable resources for SIMNET. Like the lack of a dedicated OPFOR, the lack of an Operations Group or some mechanism for providing the comparable function, significantly detracts from SIMNET's capability of serving as a CTC surrogate.

Summary

SIMNET is an important new training device in the Army's family of training devices. In fact, SIMNET and its production version the CCTT, will likely be a major component of the Army arsenal of training devices in the next decade. Yet, the strategy for using SIMNET in order to derive maximum training benefit has not been articulated. If one views the CTCs as the current pinnacle of collective training, then the degree to which SIMNET allows for a similar training experience would be a major benefit to be derived from its fielding. As indicated above, SIMNET satisfies most of the requirements to produce a CTC-like training experience. The notable deficiencies are the lack of a dedicated OPFOR and the lack of an Operations Group. Each of these can be remedied though there are potential significant cost implications in both situations.

Finally, it will also be necessary to insure that a unit performance measurement system be created for SIMNET that is comparable to that of a CTC. The importance of this topic has been addressed by Hiller (1987) in his article concerning deriving useful lessons from combat simulations including CTCs. It should be noted that ARI has funded research that will insure that such a system is developed and further that appropriate training strategies and supports are available for SIMNET when it is fielded by the Army.

References

Hiller, J. H. (1987), Designing useful lessons from combat simulations, Defense Management Journal, Second and Third Quarter, p. 29-33.

Beecher, R. G. (1987), Strategies and Standards - An evolutionary view of training devices. Fort Leavenworth, KS: Combined Arms Training Activity.

STANDARDIZATION OF TRAINING MEASUREMENT
AT COMBAT TRAINING CENTERS (CTCs)

Robert H. Sulzen
Army Research Institute - National Training Center Research Team
Fort Irwin, California

This panel describes the relationships of the U.S. Army Research Institute (ARI), the BDM Corporation, elements of the Combined Arms Training Activity (CATA) of the Combined Arms Center (CAC), and the Directorate for Army Ranges and Targets/Combat Training Centers (DART/CTC) in the conduct of an ongoing program in support of a standardized training measurement system at the combat training centers. The papers are organized around a collective training measurement research plan, current measurement and analysis systems for the CTCs in general and the National Training Center (NTC), Fort Irwin, California, and the Joint Readiness Training Center (JRTC), Fort Chaffee, Arkansas, specifically, and current methods to improve the fidelity of the training measurement system. In the context of the presentations, the discussions described the collective training measurement, instrumentation systems, and standardization plans for the NTC, JRTC, Combat Maneuver Training Center (CMTC), and Battle Command Training Program (BCTP) CTCs.

A Systems Approach to Tactical Assessment

**James T. Root
BDM Corporation**

That maneuver units must be measured as to tactical skill is self-evident. Without this assessment, there isn't any way of determining whether training is correctly tailored. There is no other method of ensuring that equipment meets current requirements. It is the only way to identify systemic issues that have been overlooked on the drawing board.

However, there is an old adage that states 'Tactics are like noses: Everyone has one'. The saying is correct. A challenge to any leader's tactical techniques is a challenge to his judgement and may result in at least one nose taking a traumatically different shape. How then do you assess maneuver units in a tactical environment? The answer is doctrine.

As an example, doctrine states that units must maintain mutual support. That is, the various elements within a unit must be able to cover each other, or at least, be able to immediately move to a location where they can cover each other. How they do it is tactics.

With this in mind, Project 2A was developed to form a doctrinally based method of assessment for maneuver units. Using the National Training Center as the laboratory, a Critical Task Measurement System (CTMS) was developed that assessed unit performance based upon how well each of various critical tasks were performed.

Briefly stated, the CTMS organized each of six identified missions that were routinely conducted at the NTC by phase (Planning, Preparation, and Execution). Each phase was then broken into sequential segments, such as Conception, Analysis, Integration, and Order for the planning phase.

At the same time, identified mission critical tasks were organized by echelon (task force, company/team, or platoon) and by battlefield operating system (Intelligence, Maneuver, Command & Control, Fire Support, Air Defense, Combat Service Support, and Mobility/Counter-Mobility).

The tasks were then sequenced and linked within the mission phases and segments. The result was a flow of inter-related tasks that continued from start to end of a particular mission.

As a quality control measure and to ensure that this system produced the data that was desired, two preliminary case studies of a Defend and a Deliberate Attack Day mission were analyzed using NTC data. A brief description of the defend mission and the functional application of 2A using one company/team is presented as results.

The Mission

The task force was ordered to Defend in Sector. To accomplish this, the task force placed two company/teams forward in the high ground facing west. Team D covered the northern approach and Team A covered the southern approach. The two remaining company/teams were placed behind them: Team C in the north and Team B in the south.

OPFOR reconnaissance elements penetrated into the task force sector during the night before the battle and were able to identify the positions of all four company/teams. Based upon this information, the OPFOR main attack, which began at dawn, followed the southern approach through the sector.

Team A fought without assistance from the rest of the task force and was quickly overwhelmed. The OPFOR assault continued against Team B.

The task force commander, seeing the battle being fought along the southern approach, ordered Team C to reinforce Team B. Although Team C moved quickly, they arrived too late to halt an OPFOR battalion from breaking through.

Team A Analysis

The following comments were extracted from the Team A Take Home Package:

Planning and Preparation

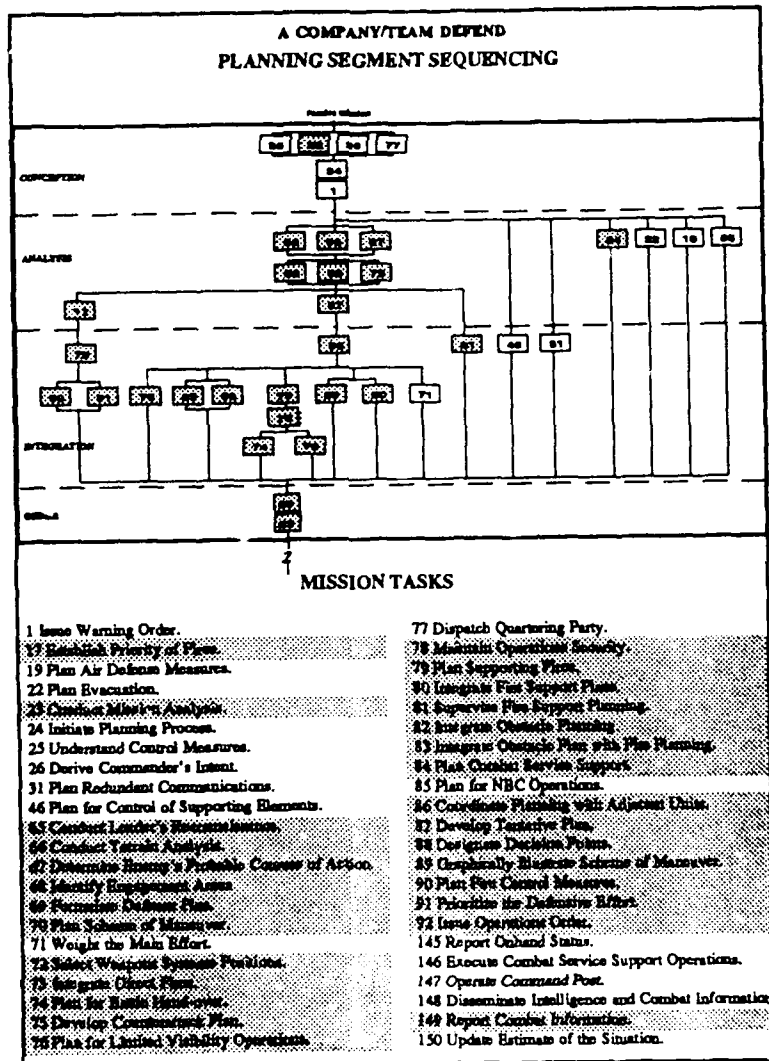
- The commander conducted a poor terrain analysis. The southern flank of the team position was completely uncovered.
- The commander did not wargame his tentative plan.
- The team failed to coordinate with forward units (in this case, the Scouts).
- The Operations Order did not address priority of work or designate rehearsals.
- The commander failed to ensure that all his subordinates understood both the plan and intent.
- The team failed to develop an adequate Fire Support plan.
- Obstacles were not addressed and no plan was developed for obstacle security and hand-over.
- The team failed to build survivable fighting positions.
- Weapons were not sited to effectively engage attacking enemy forces.
- Minefields were not marked, registered, or secured.
- Resupply was inadequate and not enough ammunition was on hand to successfully accomplish the mission.

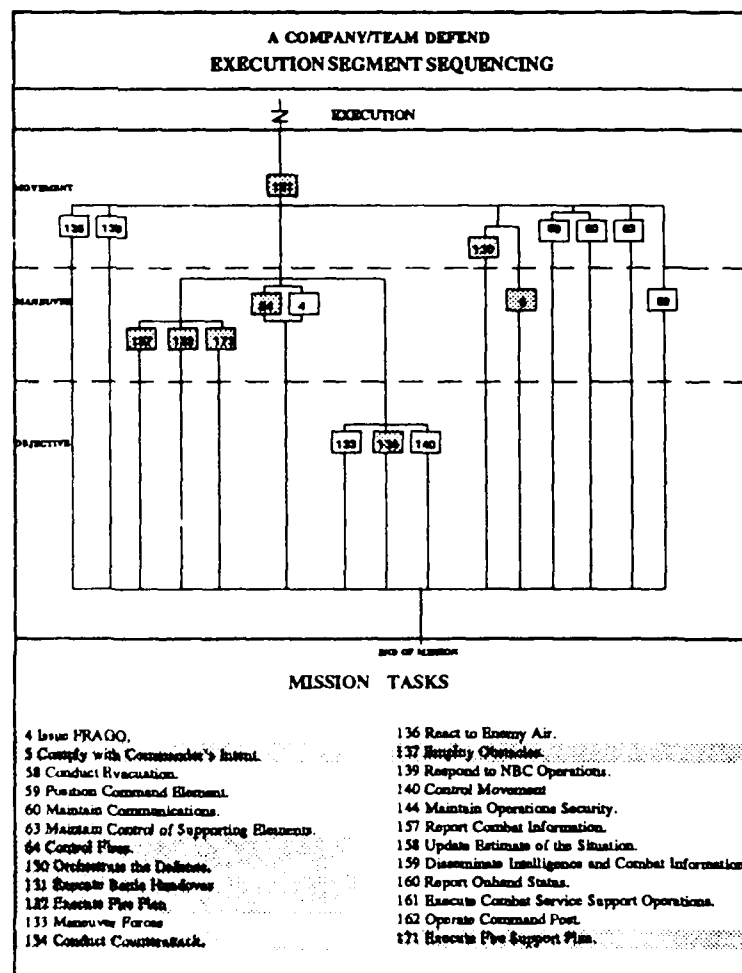
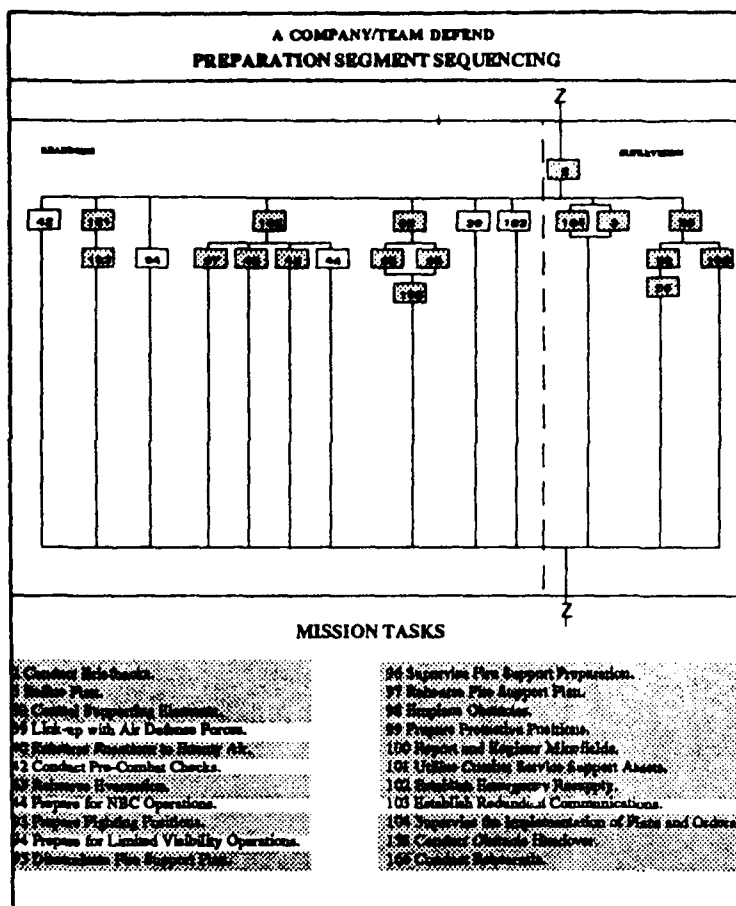
Execution

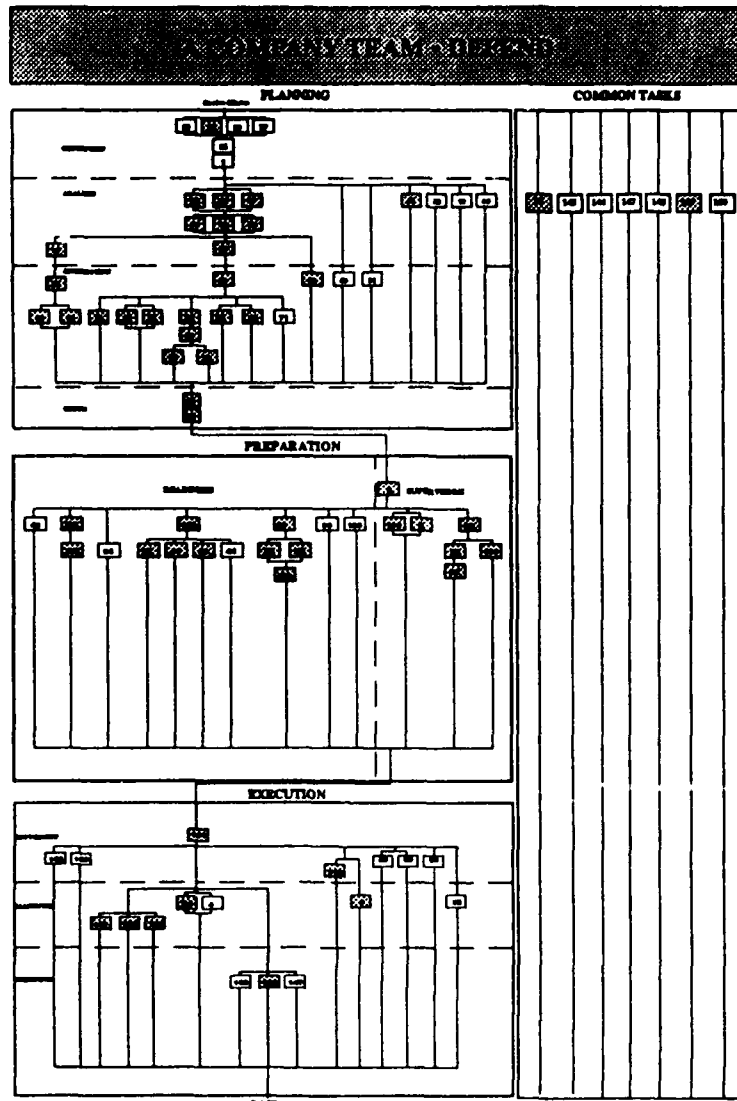
- The team lost one platoon of tanks to its own minefield.
- The team lost three BIFV's, its Aid Station, and its primary Air Defense coverage to artillery.
- The team requested only three indirect fire missions.

- Due to infrequent and inaccurate reporting, the team was unable to elicit needed support from the task force.
- Team A was wiped out and overrun in less than thirty minutes by a single OPFOR battalion.

The following Figures show an overview of the critical task analysis structure based 2A methodology. Shading indicates task deficiency. Subtasks, task elements, and measurement criteria are not shown.







Conclusion

What is immediately apparent from this analysis is why Team A was not more successful. Since the breakdowns are so clear, it is relatively easy to focus subsequent training to make appropriate corrections. Further, if similar breakdowns occur throughout the task force, then the task force has identified a collective issue. Because the system reflects the actual flow of a mission, it is possible for the assessment to be completed by the end of the mission and available as immediate training feedback to the unit.

An equally important aspect of the 2A system is the ability to rapidly analyze tactical performance for training trends and systemic issues that require resolution. Although, this is a necessarily long range effort, the system serves as a useful mechanism for data collection and organization.

In summary, 2A was developed as a tool for short term tactical training requirements and long term research. It has achieved that goal.

MEASUREMENT AT THE
JOINT READINESS TRAINING CENTER

MAJOR JOSE G. VENTURA JR.
BATTALION OPERATIONS OBSERVER CONTROLLER

"Measurement": a system of measuring. At the Joint Readiness Training Center we believe that there is, in place, an excellent system of measuring a unit's performance. Unlike NTC, JRTC does not possess the instrumentation to record a unit's movement and later play it back during the After Action Reviews (AARs). Therefore, some kind of measurement had to be developed to accomplish the same purpose.

This measurement is through the use of Training and Evaluation Outlines, better known as T&EOs and Observer Controllers. T&EOs are but a comprehensive list of tasks and the data elements that are part of these tasks. The Infantry School along with other Service Schools recently published Army Training and Evaluations Plans which contain T&EOs. Unit commanders can use them to assist in training and evaluating their units performance. All T&EOs contain the Task which is to be conducted, the conditions under which it will be conducted, and the standards which must be met if the task is to be successfully completed. Furthermore, there are numerous subtasks and standards which should be completed to increase a unit's chances of success or meeting the standards. It must be noted that these T&EOs were reviewed by the Service Schools prior to being published by JRTC. As the schools develop new T&EOs we will incorporate them where feasible.

Many will call T&EOs nothing more than checklists, however, they are much more. They are a thorough method by which to measure a unit's performance. They can be used to demonstrate the unit's strengths and those areas which need improvement. But how do these T&EOs work? Who uses them? When?

THE OBSERVER CONTROLLER TEAM

The Observer Controllers (OCs) who walk with every unit whether in the Brigade support elements or in one of the maneuver units uses a T&EO to carefully observe and later report on the unit's performance. Since there is one OC in almost every maneuver squad and at least one with every other kind of unit, it is not difficult to cover down on a unit's movement. Therefore giving us the "playback" capability required for After Action Reviews.

The JRTC Observer Controller team unlike the NTC teams is faced with controlling a very close in fight (5 meters between forces vs 50 meters at NTC). Therefore more OCs per company team are assigned to JRTC. Presently there are 12 OCs in every team. In the future it will increase to 15. This is also true for the support units, such as the Engineers and the Air Defense Artillery, more OCs are required because of the many problems encountered with the close-in fight. Figure 1 shows the current organization of the Observer Controller team.

JRTC GREENBOOKS

The OCs at the beginning of every major mission or phase are given what we call Greenbooks. These are nothing more than the T&EOs reduced to pocket size formats and protected by a green colored cover with the phase number written on it. These Greenbooks contain every T&EO which may be required by the OC. For example, during the Low Intensity Phase the following T&EOs can be used; Conduct Reconnaissance in Force, Search and Attack, Conduct Movement to Contact, Conduct a Hasty Attack, Conduct an Air Assault, etc.

Once the OC receives these Greenbooks he will read and ensure he identifies those critical and leader subtasks which need to be accomplished if a unit is to meet the task standards. A critical subtask is any task whose performance has a direct link to the accomplishment of the task. While a leader subtask can be defined as any task that must be performed by one or more leaders of the unit. The critical subtasks are clearly identified in the books by the use of a cross (+) and leader tasks use an asterisk (*).

During the conduct of the task the OCs will use the Greenbooks to answer questions pertaining to the task(s) being observed. These questions are easily answered by annotating a "yes" or "no", they either did or they did not complete the task(s). However if the event/action does not apply or it was not observed the OC will annotate by using either "N/A" or "N/O" respectively. If an OC gives the unit a "no" he must explain why in the remarks column. At the end of every T&EO the OC also has an area in which to write notes he deems necessary for the preparation of his AARs.

At the end of every phase the Greenbooks are collected and returned to the data collection section. The Greenbooks can now be used by the OCs for preparing their portions of the Take Home Package (THP) and for data collection.

T&EOs AS A DATA COLLECTION TOOL

Not only did the T&EOs create a uniform measurement they also became the primary source for data collection. The same critical and leader subtasks could also become data elements that should be routinely collected for use by such agencies as the Center for Army Lessons Learned (CALL). If a new requirement arises for data collection existing T&EOs are updated.

SUMMARY

At JRTC measurement of a unit's performance is accomplished through the use of comprehensive Training and Evaluation Outlines (T&EOs) and the Observer Controllers (OCs) who use them. These T&EOs cover every task that a unit may be required to perform. Including those tasks performed by all the Brigade Combat Support and Combat Service Support units.

OC density at JRTC is greater than at the NTC. This ensures that every movement or action the units take will be observed and later played back for the After Action Reviews (AARs). These observations are accurately recorded on the T&EOs. The OC simply records whether the action was completed or not. He accomplishes this by either checking the yes or no boxes provided in the T&EOs.

T&EOs are also used by the OCs while writing the unit's Take Home Package (THP) and for data collection. The same critical and leader subtasks are required data elements. As more data needs are identified they are quickly added to the T&EOs.

As we look to the future and bring on line instrumentation to support our data collection effort we still envision the requirement for OCs and T&EOs. However there is the possibility to use different computers (lap tops or clipboards) to accomplish what the Greenbooks now do.

The U S Army's Trendline Analysis Program

MAJ Joseph R. McLaughlin

Center For Army Lessons Learned, Ft Leavenworth, KS

Introduction. This paper describes the U S Army's Trendline Analysis Program (TLA), the name given to the collection, processing, analysis and dissemination of data produced as a result of training at the Army's four premier Combat Training Centers (CTC). These centers are the National Training Center (NTC), Ft Irwin, CA; the Joint Readiness Training Center (JRTC), Ft Chaffee, AR; the Combat Maneuver Training Center (CMTC), Hohenfels, FRG; and the Battle Command Training Program (BCTP), Ft Leavenworth, KS. The goal of this program is to provide information needed to improve doctrine, training, organization, materiel and leadership (DTOML). Additionally, standardization issues affecting this process will be discussed.

Background. The prototype CTC, the NTC, was conceived as a place that would first and foremost provide outstanding training to participating heavy battalions and secondarily provide a source of data for the identification of tactical improvements. The first mission rightly received the focus of resources and has been an unqualified success. The second, and in some respects more difficult, mission has only recently been accomplished in a comprehensive and systematic manner. The Army Research Institute - Presidio of

Monterey Field Unit (ARI-POM), sponsored by the Center for Army Lessons Learned (CALL), Combined Arms Training Activity (CATA), Ft Leavenworth, KS has recently completed a five year effort to develop unit performance measurement systems for the NTC. This effort included the development of an archive complete with the necessary procedures, personnel and equipment to support detailed exploitation of NTC data. Commanding General, U S Army Training and Doctrine Command (CG, TRADOC) verified in early 1988 the readiness of the archive to support analysis and directed his subordinate schools and centers to exploit this "goldmine" of data. As a consequence, this year saw the first widespread exploitation of NTC data and the institutionalization of procedures and facilities to ensure NTC (and objectively all CTC) data needed for analysis is collected, processed, analyzed and disseminated. The four components of the TLA process are discussed below.

Collection. Data collection to support unit training and subsequent analytical efforts to improve DTOML is a CTC mission. This mission is accomplished through a variety of automated and manual data collection systems. Collected data falls into three general categories: battlefield events (movement, firing, communications, etc.), unit performance to standard (Training and Evaluation Outline assessment), and Observer/Controller (O/C) evaluations. The quantity of data collected varies among the CTCs, reflecting the different stages of their development. The NTC, as the most mature CTC, collects a large amount of battlefield event data through its range data and communications measurement systems. These systems allow detailed analysis of task force missions down to the individual combat vehicle level. The JRTC, although not as instrumented as NTC, permits detailed task performance analysis as a result of the T&EOs employed there. BCTP, a simulation-driven CPX for divisions and corps staffs, provides simulation results and O/C assessments. The CMTC will rely initially on manual O/C data collection and assessments.

All CTCs provide the trained unit a Take Home Package (THP) consisting normally of written assessments of unit mission performance and video recordings of the mission After Action Reviews conducted by the O/Cs. This THP is provided to the unit to help them develop their training program. NTC and JRTC THPs are also forwarded to the CTC archive at ARI-POM within several weeks of the conclusion of training. BCTP data is maintained at Ft Leavenworth pending the formulation of a CATA-ARI research and development effort to exploit this operational level data. CMTC data is not yet available in significant amounts as the center transitions to initial operational capability.

CALL coordinates with the CTCs on behalf of the users to ensure required data is collected. Long-term data needs may be added to the permanent CTC collection process. Short-term data needs may be satisfied via a "focused" rotation data collection effort. This typically entails the use of checklists and surveys prepared by the proponent.

Processing. Processing arranges the data provided by the CTCs into usable forms. Data is received, cataloged, archived, and where applicable, incorporated into the relational database by ARI-POM. Many items, such as unit operations orders, THPs, AAR videos and communications tapes are simply inventoried and stored. Other data such as the raw data stream from the NTC instrumentation system is processed and may be incorporated into a relational database that contains over 300 task force missions from the NTC.

The database permits detailed analysis of battles by applying queries and on-line graphics routines. Efforts continue to expand the amount of the archive that is digitized and incorporated into the database.

Analysis. Analysis of CTC data is done routinely by CALL, the Tradoc Analysis Command (TRAC) and RAND ARROYO. TRADOC schools pursue proponent issues and ARI exploits the data in numerous research projects. Contractors have also used the archive in support of their government work.

The archive can be accessed on-site at ARI-POM or remote via post or telecommunications. Several actions must occur before access is granted. First, the user must contact CALL and receive permission to use the archive. CALL verifies the user's need for access. While archive holdings are not classified, they are highly sensitive, For Official Use Only, and exempt from the Freedom of Information Act under Category 5. CALL then coordinates with ARI-POM for user training. Quarterly week-long workshops are conducted at the archive by ARI-POM to train the users on archive procedures and provide information on the underlying CTC data collection systems. Instruction lasts one day and the remainder of the week is available for research. Once trained the user may return to conduct additional work as needed, subject to ARI-POM limits. Special training sessions can be requested through CALL.

The workshops are a convenient way to bring the users together with the CTC, CALL, TRAC and ARI representatives. These personnel provide information that allows the analyst to understand the strengths and limitations of the data collection and processing systems and hopefully preclude the analyst from being misled by the data.

Remote access via telecommunications to the mainframe hosting the database is possible for authorized users as is the mail shipment of archive holdings for temporary use.

CALL maintains the right to review user analytical results. This is done to preclude improper conclusions from being drawn from the data, verify the results of the user's classification review, and ensure that unit identities are not disclosed. CALL coordinates this review with the applicable CTC.

Dissemination. Trend information is disseminated in a number of ways. ARI and RAND publish reports and notes. CALL distributes Lessons Learned Bulletins and video tapes and conducts briefings and studies. TRADOC schools conduct studies relating to their proponentcy.

Results. Although the program did not begin to bear fruit until recently, several significant actions can be traced to

the exploitation of CTC data. The Infantry School has made major changes to its TOW training strategy. The Armor School has begun a study effort focused on improving the volume and rate of tank firing. Additionally, the Artillery School has instituted a special data collection effort at NTC to analyze ways to improve indirect fire support.

Standardization. As mentioned earlier, the CTCs have varying levels of instrumentation which affects their ability to collect data. NTC is a mature system that provides data for detailed analysis at the individual combat vehicle level. JRTC, the next most mature system, does not currently have a similar capability. This shortcoming limits analysis of light force operations as well as comparative analysis between heavy (NTC) and light operations. The instrumentation systems which provide detailed data at the NTC are mounted to the combat vehicles which are the dominant player on that battlefield. Such vehicles are limited to the opposing force at the JRTC and do not have the NTC-type instrumentation. Cheap, light and reliable instrumentation to capture dismounted soldier firing activity and movement and allow rapid retrieval for analysis would go a long way towards improving data collection gaps at the three tactical CTCs and help standardize the data collection system.

Summary. The Trendline Analysis Program is the means by which the Army collects, processes, analyzes and disseminates data produced as a result of training at the four Combat Training Centers. The goal of the program is to provide data to improve doctrine, training, organization, materiel and leadership. A key feature of the program is the CTC archive being developed by the Army Research Institute under the sponsorship of the Center for Army Lessons Learned. This facility allows for the detailed analysis of NTC battles and will eventually encompass all of the CTCs. The program enables the Army to benefit from the secondary mission of the CTCs to produce data for subsequent analysis. It has already provided important trends that have initiated proponent action.

The foregoing paper contains the opinions of the author and does not reflect an official position of the U S Army.

ARTEP MISSION TRAINING PLAN USE AT THE CTC'S

David M. Atwood
Major, U.S. Army
Collective Training Division
Combined Arms Training Activity
Fort Leavenworth, KS 66027

With the recent approval of FM 25-100 by the CSA, the Army now has a solid base on which to build its training management system. Indeed, over the past few years as FC 25-100 evolved to the version of the field manual (FM) approved in mid-October, the principles of training espoused have been inculcated in unit training programs across the Army. One of the key precepts laid out in the FM is the Battle Focus Process. Everything the unit does in training should be focused on the unit's wartime mission. The figure from FM 25-100 below is a schematic of that process.

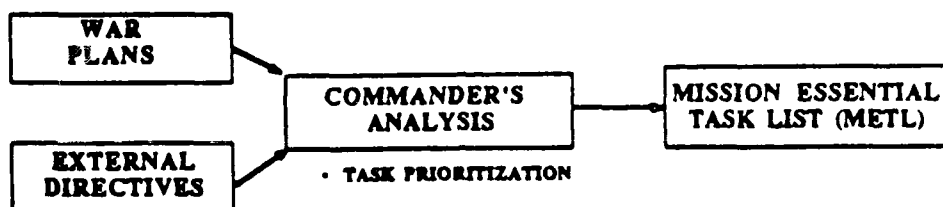


Figure 1

The METL developed by the unit commander becomes the focus of the unit's training program. Where does the MTP fit in this process? The MTP is considered a piece of the External Directives block in the diagram. It provides the unit's doctrinal missions and the associated collective tasks that comprise those missions. It also provides several matrices that tie the missions and collective tasks to exercises, references, individual tasks and leader tasks as well as its component collective tasks. Some MTPs contain a complete mission by mission hierarchy of tasks that not only relate the tasks to the mission, but also give a sequential progression of tasks accomplishment during the mission. Each type unit has or will have its proponent MTP providing a complete training package. Remember that it is a doctrinal package based on a unit's TO&E and may not include some real world missions and tasks assigned to a particular unit.

The first use of the mission training plan (MTP) at a Combat Training Center (CTC) is based on its use at homestation and its use by the commander in METL development. Prior to a unit rotation at a CTC, the commander and the operations group determine from the unit's METL those "missions" or tasks on which the unit will be evaluated during its rotation. It is a fairly simple process that ensures we gain the maximum training benefit to the unit and do not divert training resources to achieving the standard for missions or tasks that are not on the unit's METL and therefore not battle focused. In addition, figure 2 shows how the higher commander's guidance becomes the subordinate unit's METL and his own battle tasks so too does an MTP for a

achievement of the task standard but are necessary to ensure that the unit preserves its combat power while accomplishing the task and does not omit any critical steps. This ensures that luck is not a factor when the unit repeats the task at a future time.

Third, MTP usage erases most of the obscurity of the old ARTEP documents, thus tightening the subjective evaluation of the observer-controllers (OC's). With observable and measurable standards in print, all segments of the force can train to the same standard. Granted, we may have the incorrect standard for a task in an MTP, but at least we have established a reference point from which to shift until we determine the exact standard. An example of this is the use of friendly casualty figures as a measure of success. For instance, a battalion task force's mission is to defend a specific piece of terrain. In doing so, the unit should only allow 40% of its force to be wounded/killed or lose only 40% of its tanks. Although our experience at the National Training Center has shown this to be an attainable standard and that successful units have normally attained it, we have decided to stricken this type of standard from MTP's in favor of a subjective evaluator judgement that the unit preserved its force in line with the commander's intent or took all reasonable measures to evacuate wounded personnel. As you can see, we have established wider standards or at least determined that we are not ready to accept objective measurement of this type. We are, however, content with describing enemy casualty percentages as a measure of success.

Lastly, the MTP will make the training of an OC much easier by narrowing the amount of subjective judgement currently required by the CTC's and the old ARTEP documents. Doctrine has always been the basis for evaluation and training. As the detailed T&EOs are used by the OCs, their "checklists" will follow the MTP rather than be based on their interpretation of the doctrine FM, a training circular (if available) and the vague AR standards of the past. This process will not be restricted to just OCs. All users of an MTP will go through the same process and will arrive at the same standards for their respective units. There will always be variance in application of standards and certainly when the unit commander conducts his assessment, but we as a group will have a better understanding of why the variance and how the conditions have altered the exact standard. In any case, units will be able to achieve higher standards and thereby be more combat ready.

In summary, the MTP will strengthen our already viable CTC program by accomplishing these three thoughts. First, better definition of a unit's METL; second, better definition of the standards for METL tasks and lastly better training and standardization of trainers and evaluators. As shown in the figure from FM 25-100 below, units will be able to operate within the band of excellence most of the time.

unit link to its subordinate, higher or supporting MTPs in the development process. The resulting training then becomes focused on battle tasks at the CTC and also at homestation.

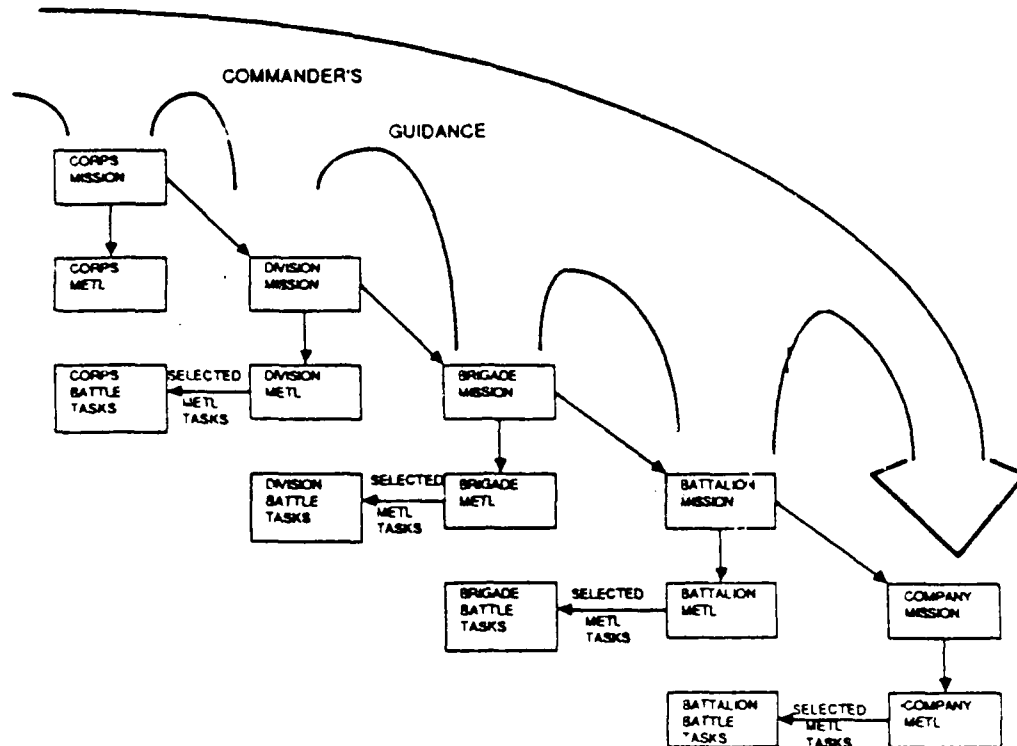


Figure 2

The second use at the CTC is derived from the Training and Evaluation Outline (T&EO's) themselves. Each T&EO also contains subtasks, conditions and standards for the collective tasks it covers. Unlike the old ARTEP documents, we have attempted to identify conditions under which each task should be performed, thereby adding validity to the published standards. As the unit trains at homestation and is evaluated at the CTC, these same conditions and standards will apply. A small point perhaps, but let's examine the details. First, each T&EO condition statement contains an OPFOR statement in the SALUTE format. What size OPFOR is needed? What is the OPFOR activity? Where is the OPFOR operating and so on. Not all SALUTE elements apply in each task but enough so that effective training of the blue force can occur. In fact, some MTPs even have OPFOR tasks or countertasks that increase the validity of the task being trained. Other condition elements are also covered but the OPFOR is key, particularly to the maneuver forces that are primary users of the CTCs. Unit commanders, by replicating the conditions as closely as possible to those in the MTP, come closer than ever before to replicating the real world battlefield and the CTC battlefield at homestation.

Second, the subtasks and standards will lead the unit through a collective task and thereby establish the correct process for achieving the task outcome standard. They do not replace

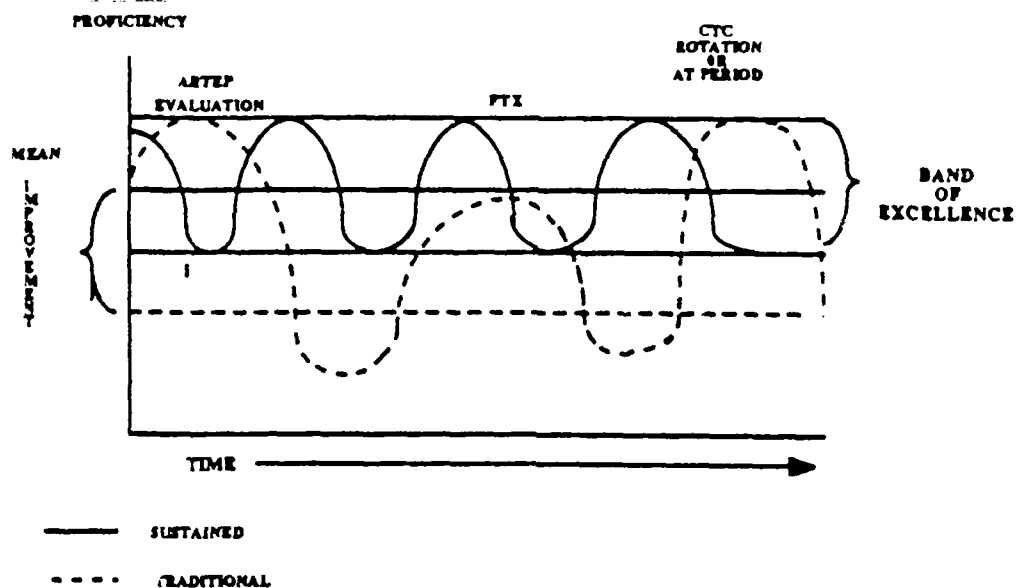


Figure 3

Rotation at a CTC will continue to be high point of unit performance but with correct application of those tasks and standards found in an MTP, a unit will be able to more closely replicate that CTC experience and battlefield at homestation and thereby lessen or eliminate the wide variances of performance based on subjective evaluation of the past.

REFERENCES:

Headquarters, Department of the Army. 15 November 1988.
Training the Force, Field Manual 25-100, Washington, D.C.

Lewman, T.J. (1987) Mission Critical Tasks at the National Training Center Research Product BDM/ARI-TR 0017-87).
 Alexandria, VA. U.S. Army Research Institute for Social and Behavioral Sciences.

DEVELOPMENT OF OBSERVER CONTROLLER (O/C) GUIDEBOOKS

Robert H. Sulzen¹

Army Research Institute-National Training Center Research Team
Ft Irwin, CA 92310-5000

Observer/Controllers

The Observer/Controllers (O/Cs) at the National Training Center (NTC) are critical to the success of the simulated battles fought on the ground at Fort Irwin, California. Imagine the Chicago Bears playing the Los Angeles Raiders without any umpires. The result might be entertaining, but it would not be a football game.

Trying to play simulated combat at the NTC without O/Cs might also be entertaining for some, but it would not be a very realistic simulation nor a valuable training experience. The O/Cs have the most vital role in providing effective simulated combat for Army tactical units.

The means of portraying the reality of combat in the field for Army units is MILES (or the Multiple Integrated Laser Engagement System). MILES uses eye-safe lasers attached to weapons as a means to safely simulate their fire. The firing signature is also simulated either by blank rounds or flash devices. The target effects are simulated by laser detectors worn by soldiers or installed on weapon systems, which when hit cause a buzzing sound to be emitted which can be shut off only by also turning off the target's firing laser.

Using the MILES during field exercises is similar to the gaming situations played by civilians in the "survival game" with paint pellet guns. Like those games, the success of MILES exercises rests on the effectiveness of the controllers or O/Cs. Enthusiastic players can make the game a shambles. Control must be applied to ensure a fair game. Strapping MILES on weapons does not ensure a simulated combat exercise anymore than strapping on shoulder pads ensures a football game. Certain rules must be adhered to if simulated combat is to occur.

While Army Combat Training Centers (CTCs) employ MILES, our research and observation demonstrates that most Army units do not properly employ MILES at home station (Roberts-Gray, Nichols, & Gray, 1984). The CTCs ensure that the MILES rules of engagement are properly followed during exercises, through the control exercised by the O/Cs. The exercises are carried out so well that according to the TRADDOC Commander (Thurman, 1988)

¹The views expressed in this paper are solely those of the author and do not reflect the views of the Department of the Army or the U.S. Army Research Institute.

"...the CTCs provide the best force-on-force experience short of actual combat in the world."

It is also through the O/Cs that at least part of any measurement of unit performance must occur. Measurement of collective performance in simulated combat is complex. Like all measurement systems, there is a need to consider validity and reliability.

Validity

Without the control provided by the O/Cs, the realism (or validity) of combat simulation would be considerably degraded. The rules of engagement which the O/Cs enforce require that participants be vulnerable to the opposing laser systems, and that when they fire their own laser weapons there is a reasonable weapons signature (Department of the Army, 1982).

As stated previously, the Army considers that the CTCs offer the most realistic training situation available. There is a strong case for face validity based upon this endorsement. However, the most valid training is done at the CTCs, while home station training seems to lack the enforcement of objective casualty assessment so necessary for validity (Roberts-Gray, 1984). As long as this situation persists, training preparation for NTC will be deficient, lacking the necessary validity of home station training.

Reliability

The consistency with which the NTC O/Cs enforce the rules of engagement is directly related to the reliability of any measurement of unit performance undertaken at the NTC. Observation of the O/C teams provides ample evidence that each team differs in their approach to similar problems. Differences are not, in and of themselves, detrimental. However, when the O/C teams differ in their enforcement of the rules of engagement, there is an opportunity for unreliable measurement.

Research Purpose of the Guidebooks

In order to improve the validity and reliability of measurement at the NTC and to assist the NTC staff, the Army Research Institute has developed O/C guidebooks on the procedures to be followed by the O/Cs.

There is a continuing requirement to train O/Cs due to the personnel turnover. In the Fall of 1986, the Commander of the Operations Group, NTC, requested Army Research Institute assistance in the training of O/Cs. A decision was made to develop O/C guidebooks to assist new O/Cs.

The first guidebook developed was a basic guidebook for force-on-force company and platoon level O/Cs on the mechanized

infantry trainer team (Scorpions), and the armor trainer team (Cobras). Four scientists were involved in the project; it was completed in December of 1987 and turned over to the Operations Group, NTC. The guidebook was published in reduced size for pocket use.

The live fire O/C team or Dragons then worked with the Army Research Institute staff to modify the company/platoon guidebook for use in the live fire training. As other guidebooks neared completion, the Dragons promptly provided feedback on the combat service support (CSS), intelligence, and fire support guidebooks in final form for their use.

At the present time there are a total of eight guidebooks published for use by the O/Cs at the NTC. Four of the guidebooks are for the force-on-force O/C teams (Scorpions and Cobras), and four are for the live fire O/C team (Dragons). The four guidebooks are similar for both groups and cover the following topics: company/platoon, CSS, intelligence, and fire support.

In the near future, guidebooks for mobility-counter mobility and NBC (nuclear, biological, and chemical) should be developed. Plans are in effect to develop guidebooks for the artillery O/C team (Werewolves), the forward support battalion O/C team (Gold Miners), antiarmor O/Cs, and the senior task force O/Cs.

Guidebook Development Methodology

The developmental techniques for the guidebooks have included a review of the military literature (doctrine), interviews with the O/Cs on the discharge of their duties, observation of the O/Cs performing their duties; and an iterative review of draft guidebooks by the O/Cs, with revisions by the Army Research Institute staff.

The iterative review of the draft guidebooks by the O/Cs was a time-consuming process, and required cross checking with sister O/C teams with similar requirements. Although this process required the expenditure of considerable time, it was valuable in that the results more closely represented the actual practices of the O/Cs and the O/Cs have a greater feeling of ownership from their participation.

The basic guidebook for company and platoon O/Cs includes a section on checks of the MILES equipment to insure that it is operational. The basic MILES manuals were consulted for the details, and the checks were sequenced in such a manner as to facilitate checking MILES operation (as opposed to performing MILES installation as emphasized in the manuals). Other sections in the basic guidebook focus on specific duties of O/Cs at the NTC.

The remaining guidebooks have been designed for O/Cs who observe and control specific tasks related to operating systems.

as performed by elements of the unit undergoing training at the NTC. These guidebooks have given the detailed duties of specific O/Cs and have required iterative review with the job incumbents.

The Army Research Institute will continue to update all guidebooks on a continuing basis as required. This will be done on a page by page basis, so that the next printing of the guidebook will incorporate the necessary changes.

Discussion

By working closely with the Observer/Controllers (O/Cs) at the National Training Center (NTC) on the development and revision of the O/C guidebooks, the Army Research Institute will assist in improving validity and reliability of unit performance measurement. The validity of measurement will be improved as the realism of the combat simulation is increased through the development and enforcement of new rules of engagement. The reliability of measurement will be strengthened as the procedures followed by the O/Cs are standardized and implemented.

When changes occur in O/C procedures that could affect the measurement of unit performance, the dates of the changes can be documented. This information will improve the accuracy of reporting results and longitudinal trends. The benefits for the Army Research Institute in developing the guidebooks for the NTC O/Cs are an improved unit performance measurement system with a data base for analysis and improvement of Army units.

References

Department of the Army. (1982). Tactical engagement simulation training with MILES. (Training Circular 25-6). Washington, DC: U.S. Government Printing Office.

Roberts-Gray, C., Nichols, J. J., & Gray, T. (1984). MILES integration support analysis, phase II. (Final Contract Report, Contract MDA 903-78-C-2014). Alexandria, VA: Army Research Institute.

Thurman, Maxwell R. (1988). "Training to Fight and Win Now and Beyond the Year 2000," Army, 38, 10. 82-85, 88-89, 91. 93, 95, 97.

PROTOTYPE ELECTRONIC CLIPBOARD SOFTWARE AND HARDWARE FOR THE COMBAT TRAINING CENTERS (CTCs)

Patrick J. Whitmarsh¹

Army Research Institute-National Training Center Research Team
Ft. Irwin, CA 92610-5000

BACKGROUND

Historically, Army Research Institute (ARI) has developed methodologies to collect valid and reliable field data. Since a systematic collection of large data sets is not currently implemented at the Combat Training Centers (CTC), valid and reliable data is problematic. With support from the Combined Arms Training Activity (CATA) of TRADOC, ARI has pursued the implementation of a hand-held computer as a methodology for the systematic collection of large sets of field data in support of research on a Unit Performance Measurement System (UPMS) at the CTCs (Atwood, Hiller, & Herman, 1986; Goehring, 1987). Based on a demonstration by ARI to GEN Thurman, the Directorate for Army Ranges and Targets/Combat Training Centers (DART/CTC) of the Army Training Support Center (ATSC), implemented a program for the CTCs to develop software and hardware for a hand-held computer known as the electronic clipboard.

This paper will present a model of the software and hardware, discuss the initial work, present results of field tryouts at the National Training Center (NTC) and the Joint Readiness Training Center (JRTC), and discuss the current work and status of the electronic clipboard.

SOFTWARE/HARDWARE MODEL

The current plan for instrumentation at the CTCs and home station include the Sun workstation and the Integrated Relational System (INGRES) data base management system (DBMS), running under the Unix operating system. Accordingly, the software model includes the application program downloaded from the application data base of INGRES DBMS on the SUN to the electronic clipboard; after data collection, the data files are uploaded to the INGRES DBMS; upon request, reports are generated through the INGRES report generator software by merging the reports data base and the data files. In addition, the hardware model is a device which is based on the software requirements, portable, rugged, and contains a RS-232 port.

¹The views expressed in this paper are solely those of the author and do not reflect the views of the Department of the Army or the U.S. Army Research Institute.

INITIAL WORK

Although the SUN workstation with INGRES DBMS was not initially available, DART/CTC conducted a front end task analysis at the JRTC and the NTC to provide the basis for the hardware and software requirements document. From the software requirements document, the application program files with attendant data files and the report formats data files were written in dBase III programming language and compiled with Clipper. The application program is presented to the user through a series of menu driven screens. Upon turning on the hardware, a logon screen is displayed. The second screen allows the entry of information about the unit to be observed. Upon exit of the second screen, the main menu appears with 10 selections available to the user. Notable is the selection of an Army Training and Evaluation Program-Mission Training Plan (A-MTP) task as a function of echelon, mission, operating system and Battle Phase (plan, prepare, execute). Each A-MTP task provides the standards for scoring by the user. In addition to scoring, the user may enter additional information for each standard through a window capability.

Depending on the format, the reports indicate the data to include comments entered by the user for each task and task standard. To represent the range of hardware requirements and software requirements necessary for the field tryouts at the NTC and JRTC, DART/CTC selected a System Research and Development Corporation (SRDC), a GridCase Plus, and a Paravant portable computers. In lieu of a 5 1/4" floppy disk drive, the Paravant utilizes a 1.2 megabyte (mb) random access memory (RAM) disk and in lieu of a keyboard, the SRDC utilizes a touch screen. Other than these notable differences and from the standpoint of the user, other dissimilarities between devices did not effect operations. Based on the initial work, DART/CTC presented demonstrations to agencies within CATA, NTC, and JRTC. As part of the demonstration, the application software was downloaded from a Zenith 184 laptop computer to the Paravant hand-held computer; data was entered into the Paravant; then the data file was uploaded to the Zenith 184 to a dBase III data base; finally, with the report generator of dBase III on the Zenith 184, a hardcopy report was generated on the printer.

FIELD TRYOUT/RESULTS

During July and August 1989 DART/CTC conducted field tryouts of the software and hardware at the NTC and JRTC. There were three SRDC, GridCase Plus, and Paravant portable computers for a total of 9 hardware devices available for the field tryouts. From selected attack and defend missions, the application software contained A-MTP tasks and standards in A-MTP format. After training on the hardware and software, each of nine O/Cs operated each of the three hardware devices during a given training mission. After each training mission each O/C responded to a questionnaire on the hardware and software. For the software, the overall results indicated including a notepad,

checklists, and individual soldier manual tasks. For the reports, the overall results indicated including reports generated in After Action Review (AAR) format, Take Home Package (THP) format; and for the checklist data, custom formats. For the hardware, the overall results indicated ensuring the device meets Mil Spec 180F (ruggedized), weighs between four to six pounds, incorporates a RAM disk (eliminate disk drives), includes an illuminated (night operations) QWERTY keyboard, incorporates a 25 X 80 screen, incorporates a tactical power supply, and mounts in the vehicle.

CURRENT WORK

Based on the field tryout results and input from ARI and CATA, Center for Army Lessons Learned (CALL), Collective Training Centers-Program Directorate (CTC-PD), and Combat Arms Training Integration Directorate (CTI); and with a release date of April 1989, DART/CTC is writing in dBase III a new version of the application software for the CTCs, developing a master A-MTP data base in INGRES-PC DBMS for the CTCs, developing the report formats for the AAR, THP, and selected checklists in INGRES-PC DBMS for NTC and JRTC, and loading specific A-MTP tasks and standards for the JRTC. ARI, CATA, and DART/CTC are formulating the plans to convert the work in INGRES-PC to INGRES on the SUN workstation with the expectation of performance prior to April 1989. For the hardware, DART/CTC is writing a Statement of Work (SOW) with the Request For Proposals (RFP) due in the Commerce Business Daily (CBD) the second quarter of FY89 with contractor performance expected in July/August 1989. Purchase of the electronic clipboard by the CTCs is anticipated for the 4th quarter of FY 89. With the purchase of the electronic clipboard by the CTCs, ARI will be in a position to systematically collect the large sets of field data required to validate the A-MTP tasks and task standards as part of the research on the Unit Performance Measurement System.

SUMMARY

This paper presented a software and hardware model for an electronic clipboard implemented in the context of the instrumentation system architecture at the CTCs and home station. Initially, DART/CTC developed the application software in dBase III to run on SRDC, GridCase Plus, and Paravant portable computers and the report generation software in dBase III to run on a Zenith 184 laptop computer. Based on field tryout results and input from ARI and CATA agencies, DART/CTC is writing a new version of the application software in dBase III, developing a master A-MTP data base in INGRES-PC DBMS, developing specific report formats in INGRES-PC DBMS; and letting a contract for hardware in the 2nd quarter of FY 89. Conversion from INGRES-PC to INGRES on the SUN workstation is expected prior to April 1989. With the purchase of the electronic clipboard in the 4th quarter of FY 89 by the CTCs, ARI will be in a position to systematically collect large sets of field data in support of the UPMS.

References

Atwood, N.K., Hiller, J.H. & Herman, J. (November, 1986). The Electronic Clipboard: A Central Requirement for Effective Automation of Training in Managment in Military Units. Proceedings of the 28th Military Testing Association Conference. (pp. 242-247) Mystic, CT: Military Testing Association.

Goehring, D.J. (October, 1987). Automated Data Capture: Electronic Clipboard System. Proceedings of the 29th Military Testing Association Conference. (pp.463-468) Ottawa, Ontario, Canada: Military Testing Association.

MILITARY TESTING ASSOCIATION

1988 ANNUAL CONFERENCE

STEERING COMMITTEE MEETING

28 NOVEMBER 1988

The meeting of the Steering Committee for the 30th Annual Conference of the Military Testing Association was held in the Jackson Room of the National Clarion Hotel, Arlington, Virginia, on Monday, 28 November 1988.

MEMBERS AND ATTENDEES: Please see "List of Members and Attendees" following these minutes. The Defense Activity for Non-Traditional Education Support and the National Headquarters for Selective Service were not represented.

ITEM

REMARKS

I. OPENING REMARKS

1. The President, Colonel Jon W. Blades, welcomed the members and opened the meeting at 1000 hours.

II. REVIEW OF 1987 MINUTES

2. The minutes of the 1987 Steering Committee were reviewed.

MOTION: It was moved by Captain Ed Naro that the minutes of the 1987 Steering Committee meeting be adopted.

SECONDED BY: Mr Charlie Holman

CARRIED

III. FINANCE

3. Mr. Holman reported that the funds received from the Canadian Forces were in excess of \$7500; these funds were used as "seed" money for this year's conference.

IV. PROGRAM

4. Dr. Art Gilbert reported that 165 abstracts for individual or panel presentation were received and that four of these were withdrawn.

V. CHANGE TO BY-LAWS

5. Mr. Holman reported on the proposed revision to the MTA By-Laws. The proposed change would limit membership in the Association to those in agencies of the associated armed forces or other governmental agencies. A new category of participants would be created designated

ITEM

REMARKS

as non-member participants. This proposal calls for deleting the phrase educational, business, industrial and private in Article 111. B.1 and Article V. B. of the existing Bylaws. Also, it calls for an addition to Article III of the By-Laws as follows:

"C. Non-Member Participants.

1. Non-member participants may participate in the annual conference, present papers and participate in symposium/panel sessions. Non-members will not attend the meeting of the Steering Committee not have a vote in association affairs."

Mr. Holman pointed out that this revision is necessary to conform to the U.S. Department of Defence and each service regulation which prohibits military and civilian personnel from accepting offices in associations based on an official military position or assignment if that association has members in it which could be providing goods or services to Department of Defence. Mr. Holman reported that Steering Committee voted for these changes in a mail ballot and that a vote of the general membership would be called for in the afternoon.

NOTE: The proposed changes to the Bylaws were approved by a vote at the General Membership Meeting held at 1530 hours on 28 November 1988.

VI. HARRY GREER AWARD

6. Dr. Gilbert announced that there was one nomination for this award and that the nominee was approved by a majority in a mail ballot. Colonel Blades is to announce the recipient on Thursday evening.

VII. MTA MEMBERSHIP

7. Dr. Gilbert proposed that the Directorate of Personnel Selection, Research and Second Careers of the National Defence Headquarters of the Canadian Forces be admitted to membership in the Association and he gave a brief sketch of their contributions to the Association.

MOTION: It was moved by Dr. John Ellis that the Directorate of Personnel Selection, Research and Second Careers, National Defence Headquarters, Canadian Forces, be accepted as an institutional member of MTA.

SECONDED BY: Richard Lanterman

CARRIED

VIII. NATIONAL HEADQUARTERS SELECTIVE SERVICE SYSTEM

8. Dr. Gilbert suggested that a letter be sent to the National Headquarters Selective Service System to terminate institutional membership of that organization in MTA because of lack of

ITEM

REMARKS

participation. Dr. Martin Wiskoff suggested an alternative proposal which would provide for a letter be written to any institutional member organization which fails to have representation on the Steering Committee for two consecutive years asking if it wished to continue its membership.

MOTION: It was moved by Dr. Peg Smith that a letter be sent to any institutional organization which failed to have representation on the Steering Committee for two consecutive years asking that organization if it wished to continue its membership.

SECONDED BY: Commander Fred Wilson

IX. U.S. COAST GUARD REPRESENTATION ON THE STEERING COMMITTEE

9. Mr. Lanterman announced that Headquarters U.S. Coast Guard would be the Coast Guard representative on the MTA Steering Committee in place of the U.S. Coast Guard Institute.

CARRIED

X. MTA CONFERENCE IN THE FEDERAL REPUBLIC OF GERMANY

10. Dr. Frederick Steege remarked on the willingness of the German Federal Ministry of Defense to host a Conference in 1991. He said that it probably would be held in Munich which was the site of the 1984 Conference. Dr. Steege expressed the feeling that the good attendance by delegates by delegates from the United States was due to a letter written at United States Department of Defense level. Dr. Steege suggested that a similar letter would be useful in support of the 1991 Conference. The discussion which followed Dr. Steege remarks suggested the feasibility of obtaining such a letter.

XI. 1989 MEETING

11. Colonel Blades commented that the 1989 convention will be held in San Antonio, Texas, 6 - 10 November 1989 and that it will be coordinated by the Air Force Human Resources Laboratory and the USAF Occupational Measurement Center.

XII. OTHER FUTURE MEETINGS

12. It was agreed that other future next two meetings would be as follows:

- a. 1990 - Pensacola, Florida, coordinated by the U.S. Navy Education and Training Program Management Support Activity
- b. 1991 - Munich, Federal Republic of Germany, coordinated by the Federal Ministry of Defense

ITEM

REMARKS

XIII. ADJOURNMENT

13. There being no further business, the meeting was meeting was adjourned at 1030 hours.

MILITARY TESTING ASSOCIATION

STEERING COMMITTEE MEETING

28 NOVEMBER 1988

LIST OF MEMBERS AND ATTENDEES

Mike Berger
Veterans Administration
Washington, DC 20420
(202) 233-6232

COL Jon W. Blades
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600
(202) 274-3265

Eugene F. Burke
Science 3 (Air)
Ministry of Defence
Lacon House, Theobalds Road
London WC1X 8RY
England
430-6335/6179

Lloyd Burtch
Human Resources Laboratory
Brooks AFT TX 78235
AV 340-3611

Dr. Heinz J. Ebenrett
GAF Administration Office
Bonn
Federal Republic of Germany
49/228/122074

Dr. John A. Ellis
Navy Personnel R&D Center
Code 15
San Diego, CA 92152-6800

Dr. Arthur C. F. Gilbert
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600
(202) 74-3265

Jeff Higgs
NDHQ/DMOS
Ottawa, Ontario K1A 0K2
Canada
(613) 992-7069

Charlie Holman
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600
(202) 274-3266

Tracye Julien
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600
(202) 274-3265

Richard S. Lanterman
U.S. Coast Guard
G-PWP
Washington, DC 20593

Mike Lentz
NETPMSA
Code 301
Pensacola, FL 32509
(904) 452-1369

Captain Francois Lescreve
CRS (Sec Psy Ond)
BRUYNSTRAA"
B-1120 Brussel (NOH)
Belgium
32/2 268 0050 Ext 3219/3319

Suzanne Lipscomb
Air Forces Human Resources Laboratory
AFHRL/PRP
San Antonio, TX 78235
AV 240-3853

1Lt Kanthleen M. Longmire
AFHRL/MOD
Brooks AFB TX 78232
(512) 536-3648

CAPT Edward Naro
NODAC
Washington Navy Yard Anacostia
Bldg 150
Washington, DC 20374

Colonel Frank Pinch
Director, Personnel Selection
Research & Second Careers
Ottawa, Ontario K1A 0K2
Canada
(613) 992-0244

W. A. Sands
Navy Personnel R&D Center
Test Systems Dept (Code 13)
San Diego, CA 952152-6800
(619) 553-9266

Lt Col Lawrence Short
AFHRL/MOD
Brooks AFB TX 78232
(512) 536-3648

Dr. Peg Smith
NETPMSA
Pensacola, FL 32509
(904) 452-1369

Dr. Friedrich W. Steege
FMOD
Bonn
Federal Republic of Germany
49/228/128242

J. S. Tartell
USAF OMC
Randolph AFB TX 78150
AV 487-6623

Capt(N) Mark H.D. Taylor
ADM (PER)/CPCSA/DMOS
NDHQ Ottawa
Ottawa, Ontario K1A 0K2
Canada
(613) 996-4580

Dr. Janet Treichel
NODAC
Washington Navy Yard Anacostia
Bldg 150
Washington, DC 20374

Commander Fred Wilson
Canadian Forces
Personnel Applied Research Unit
Suite 600
4900 Yonge Street
Willowdale, Ontario M2N 6B7
Canada
(416) 224-4964

Dr. Martin F. Wiskoff
Defense Personnel Security
Research & Education Center
Monterey, CA 93940

HARRY H. GREER AWARD

One nomination for the Harry H. Greer Award was received this year. The nominee, Mr. Walter Birdsall, as adjudged to possess those qualities required to deserve recognition for his contribution to the Military Testing Association.

A plaque was presented to Mr. Birdsall on 1 December 1988 at the annual banquet and dance of the Association. A copy of the citation on the plaque presented to Mr. Birdsall is shown on page the page 935.

MILITARY TESTING ASSOCIATION

1988

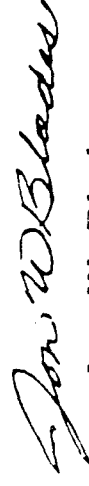
HARRY H. GREER AWARD

TO

WALTER BIRDSALL

For his contributions to the objectives and growth of the Military Testing Association from its inception. His participation, leadership, and encouragement provided impetus for the continued growth of the organization especially during the critical early years. His professionalism and expertise have contributed significantly to the professional stature of the Military Testing Association which it enjoys on this 30th anniversary.

December 1, 1988



Jon W. Blades
Colonel, Infantry
President

BY-LAWS OF THE MILITARY TESTING ASSOCIATION

Article I - Name

The name of this organization shall be the Military Testing Association.

Article II - Purpose

The purpose of this Association shall be to:

A. Assemble representatives of the various armed services of the United States and such other nations as might request to discuss and exchange ideas concerning assessment of military personnel.

B. Review, study, and discuss the mission, organization, operations, and research activities of the various associated organizations engaged in military personnel assessment.

C. Foster improved personnel assessment through exploration and presentation of new techniques and procedures for behavioral measurement, occupational analysis, manpower analysis, simulation models, training programs, selection methodology, survey and feedback systems.

D. Promote cooperation in the exchange of assessment procedures, techniques and instruments.

E. Promote the assessment of military personnel as a scientific adjunct to modern military personnel management within the military and professional communities.

Article III - Participation

The following categories shall constitute membership within the MTA:

A. Primary Membership.

1. All active duty military and civilian personnel permanently assigned to an agency of the associated armed services having primary responsibility for assessment for personnel systems.

2. All civilian and active duty military personnel permanently assigned to an organization exercising direct command over an agency of the associated armed services holding primary responsibility for assessment of military personnel.

B. Associate Membership.

1. Membership in this category will be extended to permanent personnel of governmental organizations engaged in activities that parallel those of the primary membership. Associate members shall be entitled to all privileges of primary members with the exception of membership on the Steering Committee. This restriction may be waived by the majority vote of the Steering Committee.

C. Non-Member Participants.

1. Non-members may participate in the annual conference, present papers and participate in symposium/panel sessions. Non-members will not attend the meeting of the Steering Committee nor have a vote in association affairs.

Article IV - Dues

No annual dues shall be levied against the participants.

Article V - Steering Committee

A. The governing body of the Association shall be the Steering Committee. The Steering Committee shall consist of voting and non-voting members. Voting members are primary members of the Steering Committee. Primary membership shall include:

1. The Commanding Officers of the respective agencies of the armed services exercising responsibility for personnel assessment programs.

2. The ranking civilian professional employees of the respective agencies of the armed service exercising primary responsibility for the conduct of personnel assessment systems.

3. Each agency shall have no more than two (2) representatives.

B. Associate membership of the Steering Committee shall be extended by majority vote of the committee to representatives of various governmental organizations whose purposes parallel those of the Association.

C. The Chairman of the Steering Committee shall be appointed by the President of the Association. The term of office shall be one year and shall begin the last day of the annual conference.

D. The Steering Committee shall have general supervision over the affairs of the Association and shall have the responsibility for all activities of the Association. The Steering Committee shall conduct the business of the Association in the interim between annual conferences of the Association by such means of communication as deemed appropriate by the President or Chairman.

E. Meeting of the Steering Committee shall be held during the annual conferences of the Association and at such times as requested by the President of

the Association or the Chairman of the Steering Committee. Representation from the majority of the organizations of the Steering Committee shall constitute a quorum.

Article VI - Officers

A. The officers of the Association shall consist of a President, Chairman of the Steering Committee and a Secretary.

B. The President of the Association shall be the Commanding Officer of the armed services agency coordinating the annual conference of the Association. The term of the President shall begin at the close of the annual conference of the Association and shall expire at the close of the next annual conference.

C. It shall be the duty of the President to organize and coordinate the annual conference of the Association held during his term of office, and to perform the customary duties of a president.

D. The Secretary of the Association shall be filled through appointment by the President of the Association. The term of office of the Secretary shall be the same as that of the President.

E. It shall be the duty of the Secretary of the Association to keep the records of the association, and the Steering Committee, and to conduct official correspondence of the association, and to insure notices for conferences. The Secretary shall solicit nominations for the Harry Greer award prior to the annual conference. The Secretary shall also perform such additional duties and take such additional responsibilities as the President may delegate to him.

Article VII - Meetings

A. The Association shall hold a conference annually.

B. The annual conference of the Association shall be coordinated by the agencies of the associated armed services exercising primary responsibility for military personnel assessment. The coordinating agencies and the order of rotation will be determined annually by the Steering Committee. The coordinating agencies for at least the following three years will be announced at the annual meeting.

C. The annual conference of the Association shall be held at a time and place determined by the coordinating agency. The membership of the Association shall be informed at the annual conference of the place at which the following annual conference will be held. The coordinating agency shall inform the Steering Committee of the time of the annual conference not less than six (6) months prior to the conference.

D. The coordinating agency shall exercise planning and supervision over the program of the annual conference. Final selection of program content shall be

the responsibility of the coordinating organization.

E. Any other organization desiring to coordinate the conference may submit a formal request to the Chairman of the Steering Committee, no later than 18 months prior to the date they wish to serve as host.

Article VIII - Committees

A. Standing committees may be named from time to time, as required, by vote of the Steering Committee. The chairman of each standing committee shall be appointed by the Chairman of the Steering Committee. Members of standing committees shall be appointed by the Chairman of the Steering Committee in consultation with the Chairman of the committee in question. Chairmen and committee members shall serve in their appointed capacities at the discretion of the Chairman of the Steering Committee. The Chairman of the Steering Committee shall be ex officio member of all standing committees.

B. The President, with the counsel and approval of the Steering Committee, may appoint such ad hoc committees as are needed from time to time. An ad hoc committee shall serve until its assigned task is completed or for the length of time specified by the President in consultation with the Steering Committee.

C. All standing committees shall clear their general plans of action and new policies through the Steering Committee, and no committee or committee chairman shall enter into relationships or activities with persons or groups outside of the Association that extend beyond the approved general plan of work without the specific authorization of the Steering Committee.

D. In the interest of continuity, if any officer or member has any duty elected or appointed placed on him, and is unable to perform the designated duty, he should decline and notify at once the officers of the Association that he cannot accept or continue said duty.

Article IX - Amendments

A. Amendments of these By-Laws may be made at any annual conference of the Association.

B. Amendments of the By-Laws may be made by majority vote of the assembled membership of the Association provided that the proposed amendments shall have been approved by a majority vote of the Steering Committee.

C. Proposed amendments not approved by a majority vote of the Steering Committee shall require a two-third's vote of the assembled membership of the Association.

Article X - Voting

All members in attendance shall be voting members.

Article XI - Harry H. Greer Award

A. Selection Procedures:

1. Recipients of the Harry H. Greer Award will be selected by a committee drawn from the agencies represented on the MTA Steering Committee. The CO of each agency will designate one person from that agency to serve on the Awards Committee. Each committee member will have attended at least three previous MTA meetings. The member from the coordinating agency will serve as chairman of the committee.

2. Nominations for the award in a given year will be submitted in writing to the Awards Committee Chairman by 1 July of that year.

3. The Chairman of the committee is responsible for canvassing the other committee members to arrive at consensus on the selection of a recipient of the award.

4. No more than one person is to receive the award each year, but the award need not be made each year. The Awards Committee may decide not to select a recipient in any given year.

5. The annual selection of the person to receive the award, or the decision not to make an award that year, is to be made at least six weeks prior to the date of the annual MTA Conference.

B. Selection Criteria:

The recipients of the Harry H. Greer Award are to be selected on the basis of outstanding work contributing significantly to the MTA.

C. The Award:

The Harry H. Greer Award is to be a certificate normally presented to the recipient during the Annual MTA Conference. The awards committee is responsible for preparing the text of the certificate. The coordinating agency is responsible for printing and awarding the certificate.

Article XII - Fractment

These Bylaws shall be in force immediately upon acceptance by a majority of the assembled membership of the Association and/or amended (in force 28 November 1988).

AGENCIES REPRESENTED BY MEMBERSHIP ON THE STEERING MIA COMMITTEE

Belgian Armed Forces Psychological Research Section
Canadian Forces Directorate of Military Occupational Structures
Canadian Forces Directorate of Personnel Selection, Research and Second Careers
Canadian Forces Personnel Applied Research Unit
Defense Activity for Non-Traditional Education Support
Federal Republic of Germany Ministry of Defense
National Headquarters Selective Service System
Royal Australian Air Force Evaluation Division
U.K. Science 3 Air
U.S. Air Force Human Resources Laboratory
U.S. Air Force Occupational Measurement Center
U.S. Army Research Institute
U.S. Coast Guard
U.S. Defense Personnel Security Research and Education Center
U.S. Naval Education and Training Program Development Center
U.S. Navy Occupational Data Analysis Center
U.S. Navy Personnel Research and Development Center

(as of 28 November 1980)

CONFERENCE REGISTRANTS

LT Robert R. Albright
U.S. Coast Guard Academy
New London, CT 06320

Dr. Karla Eve Allan
The BDM Corporation
2600 Garden Road, North Bldg.
Monterey, CA 93940

Deborah Allen
College for Human Services
345 Hudson Street
New York, NY 10014

John P. Allen
2009 Carriage Court
Vienna, VA 22180

Laurel E. Allender
U.S. Army Research Institute
P. O. Box 6057
Fort Bliss, TX 79916

Dr. William E. Alley
AFHRL
Brooks AFB, TX 78235

Jane M. Arabian, Ph.D.
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

MAJ David M. Atwood
11 Walker Avenue
Fort Leavenworth, KS 66027

William T. Badey
SSC-NCR
ATTN: ATNC-MOT
200 Stovall Street
Alexandria, VA 22332-0400

Annette G. Baisden
Naval Aerospace Medical
Institute (Code 412)
NAS Pensacola
Pensacola, FL 32508-5600

Dr. Hebert G. Baker
NPRDC (Code 12)
San Diego, CA 92152-6800

Louis E. Banderet, Ph.D.
USARIEM
Health and Performance Division
Natick, MA 01760-5007

LCdr Edward G. Barnett
CF Fleet School Halifax
CSE Division
CFB Halifax, Nova Scotia
Canada B3K 2X0

Betty J. Batts
6502 Eric Street
Huntsville, AL 35810

Roya Bauman
1011 Arlington Boulevard #742
Arlington, VA 22209

Michael E. Benedict
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Winston R. Bennett
AFHRL/IEDT
Brooks AFB, TX 78235-5601

LCdr Dominique M. D. Benoit
NDHQ/DMOS
100 Metcalfe Street
Ottawa, Ontario K1A 0K2
Canada

Mr. B. Michael Berger
Veterans Administration
810 Vermont Avenue, N. W.
Washington, DC 20420

LT Thomas G. Berstene
U.S. Coast Guard (G-PTE-3)
2100 2nd Street, S.W.
Washington, DC 20593-0001

Major Conrad G. Bills
ASD/ENET
Wright-Patterson AFB, OH
45433-6503

Walter Birdsall
4441 Soundside Drive
Gulf Breeze, FL 32561

Andrea E. Birnbaum
5539 Columbia Pike #815
Arlington, VA 22204

Dr. Barbara A. Black
U.S. Army Research Institute
Field Unit Ft Knox
ATTN: PERI8-IK
Fort Knox, KY 40121-5620

COL Jon W. Blades
U.S. Army Research Institute
5001 Eisenhower Ave
Alexandria, VA 22333-5600

John Boeddeker
509 Rosevelt Blvd Apt D224
Falls Church, VA 22044

Dr. Stanley F. Bolin
6772 Baron Road
McLean, VA 22101

Ms. Rose Webb Brooks
Christopher Square #88
Radcliff, KY 40160

Keith A. Brothers
2500 S. Washington Avenue
Lansing, MI 48913

Lt James H. Brown
Omaha MEPS
7070 Spring Street
Omaha, NE 68106

Dr. Larry D. Brown
E & A Division
HQ Cadet Command
Fort Monroe, VA 23651

Lawrence S. Buck
Planning Research Corporation
1440 Air Rail Avenue
Virginia Beach, VA 23455

Gary R. Bunde
8237 Lyric Drive
Pensacola, FL 32514

Eugene F. Burke
Science 3 (Air)
Ministry of Defence
Lacon House, Theobald Road
London WC1X 0RY, UK

Richard L. Burse
USARIEM (SGRD-UE-AR)
Natick, MA 01760-5007

Dr. Lloyd D. Burtch
AFHRL/PR
Brooks AFB, TX 78235-5601

Dr. Jacqueline A. Caldwell
USTAPA (CPD-R)
200 Stovall Street
Alexandria, VA 22332-0310

Mr. William B. Camm
8812 Cameo Square
Springfield, VA 22152-2222

J. R. C. Campbell
1570 Verchere Street
Orleans, Ontario K1C 4G5
Canada

Dr. Jeffrey A. Cantor
CH Associates, Inc.
1126 Kersey Road
Silver Spring, MD 20902

Edmund J. Carberry
7303 Ridan Way
Louisville, KY 40214

Thomas R. Carretta, Ph.D.
AFHRL/MOEA
Brooks AFB, TX 78235

Capt Jarean L. Carson
HQ AFMPC/DPMYOT
Randolph AFB, TX 78150-5000

Captain Richard W. Chevrier
CFPARU
4900 Younge Street, Suite 600
Willowdale, Ontario M2N 6B7
Canada

Lt Col Jeremy J. Clara
Ministry of Defense (AE4 4)
Court Road, Eltham
London SE9 5NR UK

Dr. Betty H. Colletti
HSETC (NMC/NCR)
Bethesda, MD 20814-5022

Edward M. Connelly
1625 Autumnwood Drive
Reston, VA 22094

Lowell A. Cooper
2414 Old Robinhood Road
Havre De Grace, MD 21078

Stephen M. Cormier
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Bruce V. Corsino, Psy.D.
257 Jefferson Street
Winchester, VA 22601

Jennifer L. Crafts
3333 K Street
Washington, DC 20007

Engin H. Crosby
402 Harwood Road
Harwood, MD 20776

Dr. Mark Y. Czarnolewski
U.S. Army Research Institute
ATTN: PERI-RS
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

LT Dianne L. Daniels
3009 Ashley Green Court
Waldorf, MD 20602
Lt Col William A. Daniels
Defense Intelligence College
Washington, DC 20340-5485

Lt Col Mickey R. Dansby
340 Kingston Road
Satellite Beach, FL 32937

Douglas Davis, Ph.D.
Trng. & Ed. Development Center
Marine Corps Combat Dev. Cmd.
Quantico, VA 22134-5000

Hugo De Mey
Wildebrake 6
Q040 Gent, Belgium

Major Ron A. V. Dickenson
Royal Military College
of Canada (MLM Dept.)
Kingston, Ontario K7K 5L0
Canada

Dr. Grover E. Diehl
ECI/EDXV
Gunter AFB, AL 36118-5643

Maj Robert M. Donofrio
National Defense HQ
Ottawa, Ontario K1A 0K2
Canada

Dorothy L. Dorsey
Central Sector, USMEPCOM
ATTN: MEPCO-OT-T
2500 Green Bay Road
North Chicago, IL 60064-3094

Michael Drillings
2802 Key Boulevard
Arlington, VA 22201

Dr. Walter E. Driskill
6422 Falls Church
San Antonio, TX 78247

Robert S. Du Bois
Universal Energy Systems, Inc.
Fort Knox Field Office
P. O. Box 529
Fort Knox, KY 40121-0529

Capt R. Eric Duncan
HQ AFMPC/OPMYOT
Randolph AFB, TX 78150-6001

Dr. Newell Kent Eaton
U.S. Army Research Institute
ATTN: PERI-RZA
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Heinz-Jurgen Ebenrett
Federal Armed Forces Adm Off
Bonner Talweg 177
D-5300 Bonn 1
Federal Republic of Germany

George W. Elford
ETS (Suite 475)
1825 Eye Street, N.W.,
Washington, DC 20006

Timothy W. Elig
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Gary S. Elliott
P.O. Box 15
Fort Knox, KY 40121

Linda R. Elliott
AFHRL/MOAO
Brooks AFB, TX 78235-5601

Dr. John A. Ellis
Navy Pers R&D Center
(Code 51)
San Diego, CA 92152-6800

MAJ Philip J. Exner
HQMC (MA)
Washington, DC 20380-0001

Amanda J. W. Feggetter
AGI (HF), Main Building
Ministry of Defense
Whitehall
London SW1, England

Dr. Daniel B. Felker
3333 K Street
Washington, DC 20007

Theo Feng
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Fred E. Fiedler
Department of Psychology
University of Washington
Seattle, WA 98195

Dr. M. A. Fischl
HQ DA (DAPE-ZER)
The Pentagon
Washington, DC 20310-0300

Dr. Gerald P. Fisher
Booz, Allen & Hamilton, Inc.
6606 Rosecroft Place
Falls Church, VA 22043

James L. Ford
Montgomery MEPS
Building 1512
Gunter AFB, AL 36114-6638

MAJ Dr. Michael E. Freville
1913 Arboro Place
Louisville, KY 40220

Robert Frey Jr.
U.S. Coast Guard (G-FWP-2)
2100 Second Street, S. W.
Washington, DC 20593-0001

Dr. Paul A. Gade
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

CAPT Thomas J. Gallagher
3840 Davisville Road
Hatboro, PA 19040

Jay A. Gandy
U.S. OPM
Room 6462
1900 E Street, N. W.
Washington, DC 20415

Dr. Ilene F. Gast
9507 Rockport Road
Vienna, VA 22180

Dr. Sheldon H. Geller
39 Pleasant Blvd, Suite 300
Toronto, Ontario M4T 1K2
Canada

Lt Col Frank C. Gentner
ASD/ALHA
Wright-Patterson AFB
OH 45433-6503

Major Fred W. Gibson
2301 W. Newton
Seattle, WA 98199

Dr. Arthur C. F. Gilbert
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Philip D. Gillis
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Karol Girdler
The BDM Corporation
P.O. Box 550
Leavenworth, KS 66048

Dr. Lawrence A. Goldman
3335 Old Line Avenue
Laurel, MD 20707

Joe Gornick
NETPMSA
NAS Pensacola
Pensacola, FL 32509-5000

Dr. R. Bruce Gould
ArHRL/MOD
Brooks AFB, TX 78235-5601

Dr. Scott E. Graham
U.S. Army Research Institute
Fort Knox Field Unit
Fort Knox, KY 40121-5620

Kirsten T. Graham
National Computer Systems
1101 30th Street, N. W., #500
Washington, DC 20007

Dr. Michael W. Habon
Post Fach 1420, Dept. TW10
D-7990 Friedrichshafen
Federal Republic of Germany

Matthew H. Hall
AUCPD/AIS
Bldg 803
Maxwell AFB, AL 36112

James P. Hanlon
230 E. Queen Street
Chambersburg, PA 17201

Dr. Lawrence M. Hanser
13203 Moss Ranch Lane
Fairfax, VA 22033

Christine R. Hartel
U.S. Army Research Institute
ATTN: PERI-SM
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Mary Ellen Hartmann
Questar Data Systems, Inc.
2905 West Service Road
Eagan, MN 55121-1224

Melvin B. Hayden
189 St. Clair Drive
St. Simons Island, GA 31524

Capt Jeffrey A. Hays
AFHRL/OL-AK
Trailer T-1
Bergstrom AFB, TX 78743-5000

Dr. Richard D. Herring
Dynamics Research Corporation
281-MMSD
60 Concord Road
Wilmington, MA 01887-2193

Otto H. Heuckeroth
RR # 1, Box 220
Kempner, TX 76539
Jeff G. J. Higgs
NDHQ/DMOS
101 Colonel By Drive
Ottawa, Canada K1A 0K2

LT Susan J. Hill-Fiorino
NAVMILPERSCOM Det NODAC
Bldg 150
WNY Anacostia
Washington, DC 20370-1501

Dr. Jack H. Hiller
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Edward N. Hobson
11713 Lariat Lane
Oakton, VA 22124

Dr. R. Gene Hoffman
HumRRO
295 West Lincoln Trail Blvd
Radcliff, KY 40160

Major Marilyn Hoggard
94 Dorothea Drive
Dartmouth, Nova Scotia B2W 4C4
Canada

LCDR F. Douglas Holcombe
Naval Safety Center
Code 145
Norfolk, VA 23511-5796

Virginia Melissa Holland
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

CAPT Ken C. Holleman
U.S. Coast Guard (G-WP)
2100 Second Street, S. W.
Washington, DC 20593-0001

Charlie Holman
U.S. Army Research Institute
ATTN: PERI-PO
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

LCDR Michael F. Holmes
U.S. Coast Guard (G-PWP-2)
2100 Second Street, S. W.
Washington, DC 20593-0001

Dr. Robert F. Holz
U.S. Army Research Institute
ATTN: PERI-RL
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Richard L. Horst
ARD Corporation
5457 Twin Knolls Road
Columbia, MD 21045

Major Elaine Howell
HQ AFSC/PLLP
Andrews AFB, MD 20334

Lawrence D. Howell, Jr.
Automation Research Systems
4480 King Street
Suite 500
Alexandria, VA 22302

Capt Bruce D. Hyland
CF Fleet School Halifax
Halifax, Nova Scotia B3K 2X0
Canada

Joseph W. Illes
2006 Rangewood Circle
Huntsville, AL 35803

Dr. Barbara A. Jezior
U.S. Army Natick RD&E Center
STRNC-YBF
Kansas Street
Natick, MA 01760-5020

Dr. Edgar M. Johnson
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Pansy J. Johnson
Test and Measurement Section
Defense Language Institute
English Language Center
Lackland AFB, TX 78236

Karen N. Jones
HQ DA (DAPE-CP)
The Pentagon, Room 2C667
Washington, DC 20310-0300

Tracye Julie
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Axel R. Kaiser
Freiwilligenannahmestelle Mitt
Ludwig-Beck-Str 23
Postfach 300340, Dusseldorf
Federal Republic of Germany

Michael Kaplan
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Jonathan D. Kaplan, Ph.D.
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

G. G. Kay, Ph.D.
Department of Neurology
Georgetown University Hospital
3800 Reservoir Road
Washington, DC 20007

Robert S. Kennedy, Ph.D.
Essex Corporation
1040 Woodcock Road, #227
Orlando, FL 32803

Odr Robert H. Kerr
CF Fleet School Hlifax
FMO Halifax
Halifax, Nova Scotia 8V2 2B7
Canada

John J. Kessler
U.S. Army Research Institute
5001 Eisenhower Avenue
ATTN: PERI-II
Alexandria, VA 22333-5600

Dr. Melvin J. Kimmel
U.S. Army Research Institute
ATTN: PERI-RP
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Robert W. King
NETFMSA
NAS Pensacola
Pensacola, FL 32509-5000

James R. Knight
McDonnell Douglas Astronautics
14007 Fairway Court
San Antonio, TX 78217

CAPT Douglas R. Knight, MC, USN
Naval Submarine Medical
Research Laboratory
Groton, CT 06349-5900

Juergen Kulling
Schubertstrasse 3
D-3000 Hannover
Federal Republic of Germany

LOdr Rod D. Laird
Mar Com Hdqs (Attention: CPSO)
FMO Halifax
Halifax, Nova Scotia B3K 2X0
Canada

Hyder Lakhani
3213 Amberley Lane
Fairfax, VA 22031

Dr. Ted A. Lamb
AFHRL/ID
Brooks AFB, TX 78235-5601

Richard S. Lanterman
U.S. Coast Guard (G-PWP)
2100 Second St., S.W., Rm 4200
Washington, DC 20593-0001

Michael T. Laurence
1600 Wilson Blvd
Suite 400
Arlington, VA 22209

Kenneth A. Lawrence
1111 Constitution Ave N. W.
Room 1012
Washington, DC 20224

Julia A. Leaman
4032 Caribon Street
Mitchellville, MD 20716

Lisa L. Leffler
MEPS TEST/SPEC
7070 Spring Street
Omaha, NE 68106

Jerry Lehnus
1600 Wilson Blvd
Suite 400
Arlington, VA 22209

Dr. James M. Lentz
NETPMSA
(Code 301R)
Pensacola, FL 32509-5000

Jack F. Leon, M.A.
Mohawk College
Fennell Campus, P.O. Box 2034
Hamilton, Ontario L8N 3T2
Canada

Capt Francois J. M. E. Lescreve
Sec Psy Ond/CRS
Bruynstraat
B-1120 Brussels
Belgium

Dr. Richard A. Lilienthal
U.S. TAPA
DAPC-CPD-R
200 Stovall Street
Alexandria, VA 22332-0300

M. Suzanne Lipscomb
AFHRL/PRP
Brooks AFB, TX 78235

CPT Jody C. Locklear
10304 Lyris Court, S. W.
Tacoma, WA 98498

1Lt Kathleen M. Longmire
AFHRL/MODM
Brooks AFB, TX 78235

Jeffrey N. Loube
Atlantis Aerospace Corporation
1 Kenview Boulevard
Brampton, Ontario L6T 5E6
Canada

Dr. Saul M. Luria
Medical Research Laboratory
Submarine Base
Groton, CT 06349-5900

Christina M. Lynn
11808 Henry Fleet Drive
Rockville, MD 20854

Lt(N) Chris D. F. Lyon
Maritime Command Headquarters
FMO Halifax (Attention CPARC)
Halifax, Nova Scotia B3K 2X0
Canada

Odr Donald S. MacKay
Maritime Command Headquarters
FMO Halifax
Halifax, Nova Scotia B3K 2X0
Canada

Murray J. Mack
U.S. TAPA
ATIN: DAPC-CPD-R
200 Stovall Street, Rm 4S-29
Alexandria, VA 22332-0300

Douglas H. Macpherson
U.S. Army Research Institute
5001 Eisenhower Ave
Alexandria, VA 22333-5600

Dr. Fred A. Mael
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Richard E. Maisano
171 N. Columbus Street
Arlington, Virginia 22203

Arthur Marcus
U.S. Army Research Institute
5001 Eisenhower Avenue
ATTN: PERI-SM
Alexandria, VA 22333-5600

Jerrold Markowitz
US Coast Guard (G-PRF)
2100 Second Street, SW
Washington, DC 20593

Bernie Marrero, Ph.D.
17 Rob Roy
Austin, TX 78746

Dr. Phyllis P. Marson
4500 Palm Coast Parkway, S. E.
Palm Coast, FL 32037

Dr. Michael D. Matthews
Dept. of Behavioral Sciences
Drury College
90 North Benton Avenue
Springfield, MO 65802

Joyce D. Mattson
Psychological Services, Inc.
1333 H St N. W., Suite 1100
Washington, DC 20005

Paul W. Mayberry
Center for Naval Analyses
4401 Ford Avenue
Alexandria, VA 22302-0268

Capt Colin F. Mayo
HQMC (MA)
Washington, DC 20380-0001

William P. McAleer
HQSTCFA (ROK/US)
APO SF, CA 96358-0198

Dr. D. Michael McAnulty
Anacapa Sciences, Inc.
P. O. Box 489
Fort Rucker, AL 36362-5600

Christina M. McBride
10636 High Beam Court
Columbia, MD 21044

Donald E McCauley, Jr.
Office of Rsch & Development
Room 6451
Office of Personnel Management
Washington, DC 20415

LTCDR Michael A. McKenzie
3224 Yorktown Drive
Tallahassee, FL 32312

MAJ Joe McLaughlin
2900 S. 14th Street
Leavenworth, KS 66048

Dr. Herbert L. Meiselman
U.S. Army Natick R&D Center
Natick, MA 01760

Dr. Albert H. Melter
Personalstammamt Bundeswehr
Mudra-Kaserne, Kolner Str 262
D-5000 Koeln 90
Federal Republic of Germany

Maj Harold C. Mendes
College militaire
royal de Saint-Jean
Richelain, Quebec J0J 1R0
Canada

Rex R. Michel
U.S. Army Research Institute
Field Unit
P.O. Box 349
Fort Leavenworth, KS 66025

John L. Miles, Jr.
12815 Fernwood Turn
Laurel, MD 20708

Jimmy L. Mitchell, Ph.D.
McDonnell Douglas Astronautics
926 Toepperwein Road
Converse, TX 78109-2420

Dr. John A. Modrick
Honeywell, Inc.
P. O. Box 1361
Minneapolis, MN 55440

Stanley P. Morrison
Policy and Resources Division
ETD, ODCST (ATTG-I)
HQ TRADOC
Fort Monroe, VA 23651-5000

Dr. Franklin Moses
6301 North 35th Street
Arlington, VA 22213

LOdr Don Mullin
Minto Place Hotel
Laurier Street
Ottawa, Ontario
Canada

Dr. Michael D. Mumford
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332

CAPT Edward L. Naro
NMPC, NODAC
WNY Anacostia
Washington, DC 20370-1501

Marshall A. Narva
U.S. Army Research Institute
5001 Eisenhower Ave
Alexandria, VA 22333-5600

Dr. Gilbert L. Neal
U.S. Army TRADOC Analysis Cnd
WSMR ATTN: ATRC-WG
White Sands Missile Range
NM 88002

Dr. Peter F. Newton
Box 1, FANX 3
National Security Agency
Fort Meade, MD 20755-6000

Dr. Nigel R. Nicholson
ARI Field Unit
HQ TEXCOM
Fort Hood, TX 76544-5065

Ms Mary L. Norwood
112 Melody Lane
Warner Robbins, GA 31088

Dr. Lawrence H. O'Brien
Dynamics Reserach Corporation
60 Concord Street
Wilmington, MA 02114

Brian S. O'Leary
Office of Personnel Management
Room 6451
1900 E Street, N. W.
Washington, DC 20415

Dr. Frank O'Mara
Advanced Technology, Inc.
12001 Sunrise Valley Drive
Reston, VA 22091

Carolyn Davis Oates
Battelle Washington
2030 M Street, N. W.
Washington, DC 20036-3391

Lt(N) Alan C. Okros
CFPARU
Suite 600, 4900 Yonge St.
Willowdale, Ontario M2N 6B7
Canada

Dr. Laurel W. Oliver
U.S. Army Research Institute
ATTN: PERI-RL
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Darlene M. Olson
U.S. Army Research Institute
ATTN: PERI-RS
Alexandria, VA 22333-5600

Mr. David D. J. Owen
NDHQ/DMOS 3-4
101 Colonel By Drive
Ottawa, Ontario K1A 0K2
Canada

Col Donald H. Oxley
Army Sch of Tng Support
Wilton Park
Beaconsfield, Bucks HP9 2RP
England

Ok-Choon Park
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. David L. Payne
Hay Systems, Inc.
2000 M Street, NW, Suite 650
Washington, DC 20036

Major Denis J. D. P. Pelletier
QGZR (Q)
1048 rue St-Jean
Quebec (Que) G1R 1R6
Canada

Robert H. Pennington
NETPMSA
Saufley Field
Pensacola, FL 32509-5000

Dr. Ray S. Perez
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

ENS Jim E. Peschel
U.S. Coast Guard (G-PWP)
2100 Second Street, S. W.
Washington, DC 20593-0001

Morris P. Peterson
SSC-NCR
ATTN: ATNC-MOA
200 Stovall Street
Alexandria, VA 22332-1336

Allan L. Pettie, Ph.D.
609 Tam O'Shanter Boulevard
Williamsburg, VA 23185

William J. Phalen
AFHRL/MDEPM
Brooks AFB, TX 78235-5600

Don C. Phillips
Naval Education & Tng Program
Management Support Activity
Saufley Field
Pensacola, FL 32509-5000

Col Franklin C. Pinch
NDHQ/DPSRSC
101 Colonel By Drive
Ottawa, Ontario K1A 0K2
Canada

Dr. Barbara S. Plake
Buros Inst of Mental Measmts
135 Bancroft Hall
University of Nebraska-Lincoln
Lincoln, NE 68588-0348

Dr. Robert F. Priest
Ofc of Institutional Research
U. S. Military Academy
West Point, NY 10996

LCol Terry J. Prociuk
Royal Military College
of Canada
Kingston, Ontario K7K 5L0
Canada

Dr. David M. Promisel
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

LCDR Ellen M. Quisenberry
Naval Medical Command
Washington, DC 20372-5120

Capt Gregory A. Reiser
Central Flying School
Canadian Forces Base Winnipeg
Westwin, Manitoba R3R 0T0
Canada

Beatrice J. Rheinstein
OPM
1900 E Street, Room 6451
Washington, DC 20415

William M. Ritchie
NDHQ/DGPRD
101 Colonel By Drive
Ottawa, Ontario K1A 0K2
Canada

LtCol Billy J. Roberts
Instructional Mgt. School
Training and Education Center
MCCDC
Quantico, VA 22134-5050

Capt Ed J. Ronsyn
CfB Borden
Borden, Ontario LOM 1C0
Canada

Kendall L. Roose
Academic Training Dept.
Training Airwing Five
NAS Whiting Field
Milton, FL 32570

James T. Root
Bldg 106, P.O. Box 5787
Presidio of Monterey, CA 93944

Andrew M. Rose
3333 K Street, N.W.
Washington, DC 20007

Harvey Rosenbaum
2833 Terrace Drive
Chevy Chase, MD 20815

LOdr Peter J. Ross
The Royal Naval Staff College
Greenwich, London SE10 9NN
England

William, A. Ross
The BDM Corporation
P.O. Box 550
Leavenworth, KS 66048

Dr. Paul G. Rossmeissl
2000 M Street, N.W.
Washington, DC 20036

Dr. Hendrick W. Ruck
AFHRL/ID
Brooks AFB, TX 78235-5601

Michael G. Rumsey
8807 Burbank Road
Annandale, VA 22003

Dr. Craig J. Russell
IMLR
Rutgers University
New Brunswick, NJ 08903

MAJ Charles A. Salter
10 East Militia Heights
Needham, MA 02192

Michael G. Samet
ISR Corporation
6312 Variel Avenue
Woodland Hills, CA 91367

William A. Sands
Navy Personnel R&D Center
(Code 13)
San Diego, CA 92152-6800

Dr. Joel M. Savell
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Sibylle B. Schambach
Federal Armed Forces Ad Office
Bonner Talweg 177
D-5300 Bonn
Federal Republic of Germany

Christine L. Schlichting
Box 900, NSMRL
Naval Submarine Base
Groton, CT 06349

Mr. Bernard G. Schuster
9174 Broken Oak Place
Burke, VA 22015

LTC G. Rufus Sessions
Walter Reed Institute of Resch
ATTN: SGRD-UWI
Washington, DC 20307-5100

MAJ Carl E. Settles, Ph.D.
USA MEDDAC
APO SF, CA 96343-0076

SSG Fred P. Shaw
7750 Hurontario Street
Brampton, Ontario L6V 3W6
Canada

Lt Col Lawrence O. Short
AFHRL/MOD
Brooks AFB, TX 78235-5601

Barbara L. Shukitt
USARIEM
Kansas Street
Natick, MA 01760-5007

Dr. Uldi Shvern
5909 Brookview Drive
Alexandria, VA 22310

Dr. Guy L. Siebold
124 Roberts Lane #400
Alexandria, VA 22314

Dr. Zita M. Simutis
2802 Key Boulevard
Arlington, VA 22201

Alfred L, Jr. Smith
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Elizabeth P. Smith
9624 Pierrpnnt Street
Burke, VA 22015

Dr. Margaret J. Smith
NETPMSA (Code 301)
Pensacola, FL 32509-5000

Dr. Robert G. Smith
7606 Chancellor Way
Springfield, VA 22153

Dr. Edwin R. Smootz
15238 Fauboin Trail
Leander, TX 78641

Garnett L. Spearman
2328 Travis Pines
Augusta, GA 30906

Dr. Daniel L. Stabile
NAVMILPERSCOM Det NODAC
Bldg 150, WNY Anacostia
Washington, DC 20370-1501

Paul P. Stanley II
USAFOMC/OMD
Randolph AFB, TX 78150-5000

Dr. Friedrich W. Steege
FMOD-P II 4
Postfach 1328
D-5300 Bonn 1
Federal Republic of Germany

Dr. Alma G. Steinberg
U.S. Army Research Institute
ATTN: PERI-RL
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Robert W. Stephenson
AFHRL/SA
Brooks, AFB, TX 78235-5601

Jack J. Sternberg
Advanced Technology, Inc.
P.O. Box 1493
Carmel, CA 93942-1493

Dr. Nora K. Stewart
U.S. Army Research Institute
5000 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. John E. Stewart, II
12306 Wadsworth Way
Woodbridge, VA 22192-6241

Dr. LeRoy A. Stone
P.O. Box 395
Harpers Ferry, WV 25425

Dr. Lawrence J. Stricker
Educational Testing Service
Princeton, New Jersey 08541

Maj Frank Strickland
NDHQ/DGCREW
101 Colonel By Drive
Ottawa, Ontario K1A 0K2
Canada

Dr. Robert H. Sulzen
1525 Paloma Street
Barston, CA 92311

Terry Y. Takahashi
Defense Intelligence College
Washington, DC 20340-5485

Joseph S. Tartell
USAF Occupational Meas Center
Randolph AFB, TX 78150-5000

Capt(N) Mark H. D. Taylor
National Defense Headquarters
Attention: DMOS
Ottawa, Ontario KIA OK2
Canada

Dr. Joel M. Teitelbaum
Dept of Military Psychiatry
WRAIR
Washington, DC 20307-5100

LCol (Dr.) John J. Terpstra
P. O. Box 90701
2509 LS The Hague
Netherlands

Dr. Pamela M. Terry
P.O. Box 2086
Fort Benning, GA 31905

Dr. Sharon Tkacz
Allen Corporation
209 Madison Street
Alexandria, VA 22314

Capt Michael R. Toney
Instructional Mgmt School
Training and Education Center
MCCDC
Quantico, VA 22134-5050

Barbara T. Transki
NAVMILPERCEN Det NODAC
Bldg 150, WNY Anacostia
Washington, DC 20370-1501

Dr. Janet M. Treichel
NMPC, NODAC, Bldg 150
Washington Navy Yard/Anacostia
Washington, DC 20370-1501

Dr. Trueman R. Tremble
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Joseph R. Tremul
USAOMMCS
ATIN: ATSK-TI
Redstone Arsenal
AL 35897-6600

Ms Janet Tucker
4472 Silverleaf Drive
Virginia Beach, VA 23462

Dr. Paul Twohig
U.S. Army Research Institute
ATIN: PERI-RL
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. George M. Usova
4712 N. 1st Street
Arlington, VA 22203

Dr. Guus J. P. van den Elzen
Binckhorstlaan 135 - 2500 ES
The Hague, Holland

Ms Sally J. Van Nostrand
U.S. Army Concepts
Analysis Agency
8120 Woodmont Avenue
Bethesda, MD 20814-2797

David S. Vaughan
McDonnell Douglas Astronautics
E462/327/450
P.O. Box 516
St. Louis, MO 63166

Paul R. Vaughan
136A Spanish Trail
Hampton, VA 23669

MAJ Jose G. Ventura, Jr.
2119 Steeple Chase
Jacksonville, AR 72076

Naomi Verdugo
U.S. Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Richard W. Vestewig
Perceptronics
21135 Erwin Street
Woodland Hills, CA 91367

Dr. Lloyd W. Wade
Special Programs Department
Marine Corps Institute
Arlington, VA 22222-0001

Dr. Howard Wainer
ETS (21 T)
Princeton, NJ 08541

Dr. Raymond O. Waldkoetter
U.S. Army Soldier Spt Center
ATTN: ATSG-DDN (Bldg 401)
Fort Harrison, IN 46216-5700

Clarence L. Walker
Rt. 1, Box 593
Purcellville, VA 22132

Clinton B. Walker
US Army Research Institute
5001 Eisenhower Avenue
ATTN: PERI-RS
Alexandria, VA 22333-5600

Martin R. Walker
US Army TRAC-FEHN
ATTN: ATRC-B (Bldg 401B)
Fort Harrison, IN 46216-5000

Dr. Chih-yen Wang
Special Programs Department
Marine Corps Institute
Arlington, VA 22222-0001

Catherine S. Warfield
Naval Education & Trng Program
Management Support Activity
Savfley Field
Pensacola, FL 32509-5000

Dr. Brian K. Waters
HumRRO
1100 South Washington Street
Alexandria, VA 223144499

Dr. Thomas W. Watson
AFHRL/MOAE
Brooks AFB, TX 78235-5601

Johnny J. Weissmuller
10218 Lorene Labe
San Antonio, TX 78216

Dr. Leonard A. White
U.S. Army Research Institute
ATTN: PERI-RS
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Patrick J. Whitmarsh
28451 Morro Street
Barston, CA 92311

Capt Ed I. E. Wiebe
Central Flying School
Canadian Forces Base Winnipeg
Westwin, Manitoba R3R 0T0
Canada

Major Bill W. R. Wild
Canadian Forces Base Borden
Borden, Ontario LOM 1C0
Canada

Wolfgang Wildgrube
Streitkrafteamt
Box 20 50 03
D-300 Bonn 2
Federal Republic of Germany

Robert E. Wilhelm
USA Military Police School
ATTN: ATZN-MP-DE
Fort McClellan, AL 36205-5030

1Lt John E. Williams
13217 Ryden
San Antonio, TX 78233

Carol A. Williams
CNO (OP-112D1)
Washington, DC 20350

ENS Darwyn A. Wilmoth
U.S. Coast Guard (G-FWP)
2100 Second Street, S. W.
Washington, DC 20593-0001

Commander Fred P. Wilson
CFAPRU
4900 Younge Street, Suite 600
Willowdale, Ontario M2N 6B7
Canada

Laurens Wise
AIR
3333 K Street, N.W.
Washington, DC 20007

Dr. Robert A. Wisher
US Army Research Institute
5001 Eisenhower Avenue
Alexandria, VA 22333-5600

Dr. Martin F. Wiskoff
307A Mar Vista Drive
Monterey, CA 93940

John H. Wolfe
4310 Hill Street
San Diego, CA 92107-4117

Darrell A. Worstine
SSC-NCR
ATTN: ATNC-MO
200 Stovall Street
Alexandria, VA 22332-1330

Barbara A. Wright
5909 Rosedale Drive
Hyattsville, MD 20782

Winnie Y. Young
American Institute for Rsch
3333 K Street
Suite 300
Washington, DC 20007

Victor A. Zanona
Des Moines MEPS
2500 University Avenue
Des Moines, IA 50265-1480

INDEX OF AUTHORS AND PANEL MEMBERS

- Ahmed, Z. Nagin, Lt, 740
 Albright, Robert R., 204
 Allender, Laurel, Dr., 597
 Alluisi, Earl A., 686
 Arabian, Jane M., 434
 Atwater, Leanne E., Ph.D., 129
 Atwood, David M., Major, U.S. Army, 915
 Atwood, Nancy K., 859, 894
 Baisden, Annette G., 554
 Baker, Rondald C., 372
 Baker, Herbert George, Ph.D., 571
 Baldwin, Robyn, 663
 Baltzley, Dennis R., 422
 Banderet, Louis E., 76
 Banderet, L. E., Ph.D., 76, 440
 Barnett, Edward G., LCDr, CF, 510
 Bauman, Roya, 225
 Bennett, Winston R., 778
 Benoit, Dominique D., LCDr, 268
 Bills, Conrad G., Major, 740
 Bittner, Alvah C., Jr., Ph.D., 440
 Black, Barbara, Dr., xvii
 Blunt, Janet H., 274
 Bowler, Edmund C., 621
 Brett, Bryan E., Mr., 597
 Brittain, Clay V., 608
 Brooks, Rose Webb, M.A., 222
 Brown, Dianne C., 475
 Brown, James E., 740
 Buck, Lawrence S., 404
 Bunde, Gary R., 755
 Burse, Richard L., Sc.D., 63, 82
 Bush, Brian J., 737
 Butzin, Clifford A., Ph.D., 716
 Cameron, Von M., Lieutenant, 716
 Campshure, David A., 498
 Cantor, Jeffrey A., 602
 Caretta, Thomas R., Ph.D., 559
 Carson, Jarean L., Captain, 716
 Cato, Emily, 244
 Charles, Jan, 45
 Connolly, Edward, 576, 582
 Crafts, Jennifer L., 621, 627
 Crawford, Howard W., LTC, 882
 Crumley, Lloyd M., Dr., 159
 Cymerman, A., 64, 70, 82
 Czarnolewski, Mark Y., 591
 Davis, Robert J., Colonel, 538
 Davis, Phillip A., SQDNLDR, RAAF, 280
 Dickenson, R. A. V., Major, 123
 Diehl, Grover E., 669
 Dittmar, Martin J., 360, 366
 Doherty, William J., 859, 868, 894
 Dragsbaek, Heather, 515
 Driskill, Walter E., Dr., 334, 353
 Du Bois, Robert S., 33, 165
 Duncan, R. Eric, Captain, USAF, 410
 Eaton, Newell Kent, 225
 Elliott, Linda R., 469
 Ellis, John, 663
 Eschenbrenner, A. John, 803
 Fakult, Nancy J., Lt, 740
 Falleson, Jon J., Ph.D., 183, 486
 Feggetter, Amanda J. W., 698
 Feldsott, Steve, 797
 Felker, Daniel B., 615
 Fiedler, Fred E., 135, 141
 Fiorino, Susan, LT, USN, 323
 Fleishman, Edwin A., 749
 Fober, Gene W., 761
 Freville, Michael E. Ed.D., 222
 Fulco, Charles S., 22
 Furukawa, T. Paul, LTC, Ph.D., 237
 Gade, Paul A., 225
 Gandy, Jay A., 428
 Gast, Ilene F., 653
 Geller, Sheldon H., 256
 Gentner, Frank C., Lt Col, xi, 731
 Gibson, Frederick W., 141
 Girdler, Karol, 888
 Goldman, Lawrence A., Ph.D., 286
 Gould, R. Bruce, 353
 Graham, Scott E., 641
 Greene, Charles A., 51
 Greene, Lee, COL, xviii
 Guadagnoli, Mark A., 761
 Habon, Michael, Dr., 452
 Hageman, Dwight C., 565
 Hand, Darryl, 347
 Hanlon, Jim, Professor, 538
 Hanser, Lawrence M., 811
 Harden, William R., 761
 Harding, Francis D., 749
 Harnest, Charles W., 615
 Harris, Richard, xiv
 Harrison, Kent, Colonel, xviii
 Hartel, Christine, 492
 Haynes, William, 347

Hegge, Frederick W., 504
 Herring, Richard D., 3
 Heuckeroth, Otto, 521
 Higgs, G. Jeffrey, Mr., 268
 Hiller, Jack H., xviii, 838, 839, 869
 Hirsch, Edward, 515
 Hoffer, Steven W. Sgt, USAF, 542
 Hoffman, R. Gene, 498
 Hoggard, Marilyn, Maj., CF, 681
 Holcombe, F. Douglas, LCDR, 554
 Holland, Melissa, 832
 Howell, Elaine, Major, USAF, 727
 Hyland, Bruce D., Capt., CF, 681
 Irvin, Janet George, 274
 Jezior, Barbara A., 51
 Johnson, David M., 653
 Jones, A. R., Lieutenant Commander,
 Royal Navy, 692
 Kaiser, Axel R., 527
 Kaplan, Jonathan D., 27
 Kay, Gary G., Ph.D., 440
 Kelly, Dennis R., 93
 Kennedy, Robert S., Ph.D., 422, 440
 Kerins, James W., 864
 Kerner-Hoeg, Susan, 244
 Kerr, Robert H., Cdr., CF, 681
 Klein, Gary A., 153
 Knight, Douglas R., 70, 82
 Knight, J. R., 785
 Kostyla, Stan, 832
 Kuhnert, Karl W., 200
 Laabs, Gerald J., Ph.D., 571
 Lakhani, Hyder, 147
 Lal, Rashmi, 147
 Laskowski, Marlene R., Ms, 722
 Latimer, S. L., Lieutenant Commander,
 Royal Australian Navy, 692
 LaVerne, John, 177
 Lawrence, Michael T., 216
 Leaman, Julia A., 171
 Lester, Laurie S., 515
 Lett, John A., Jr., 817
 Lickteig, Carl W., 33, 165
 Locklear, Jody C., 135
 Longmire, Kathleen M., 1Lt, 262, 366
 Lubaczewski, Thomas J., 855
 Luria, S. M., 64
 Madden, Jim L., 849
 Maier, Milton H., 614
 Marrero, Bernie, Ph.D., 773
 Marson, Phyllis Peters, Dr., 744
 Matthews, Michael D., 189
 Mayberry, Paul W., 633
 McAnulty, D. Michael, 548
 McCauley, Donald E., Jr., 298
 McHenry, Jeffrey J., xv, 434
 McIntyre, Heather M., 698
 McLaughlin, Joseph R., MAJ, 911
 Melter, Albert H., 532
 Mercatante, Theresa A., 565
 Michel, Rex R., 153
 Miles, John L., Jr., 27, 591
 Mitchell, J. L., 341, 347, 791
 Modrick, John A., 39, 710
 Moise, Samuel L., Jr., 504
 Moreland, James D., 177
 Morris, Nancy, 64
 Moses, Franklin L., 686
 Mumford, Michael D., 749
 Naro, Edward L., Captain, USN, 316,
 317
 Nau, Kathy L., 767
 Neal, Gilbert L., 177
 Nichols, Judith J., 864, 882
 Nicholson, Nigel R., 15
 O'Brien, Lawrence H., 3
 O'Leary, Brian S., 298
 O'Mara, Francis E., Ph.D., 826
 Okros, Alan C., Lieutenant (N), 481
 Osteen, Mary K., 422
 Outerbridge, Alice N., 428
 Oxford, Rebecca L., Ph.D., 822
 Oxley, Donald H., Colonel, 675
 Park, Randolph K., 811
 Penn, Robert, Ph.D., 129
 Perrin, Bruce M., 785
 Peterson, Morris, Ph.D., 244
 Pettie, Allan L., 395
 Phalen, William J., 335, 341, 347,
 360, 366
 Plake, Barbara S., 390
 Plocher, Thomas A., 710
 Popper, Richard D., 515
 Potter, Earl H., 204
 Powell, Charles G., 135
 Priest, Robert F., Dr., 193
 Psotka, Joe, 832
 Quilter, Dennis, Major, 675
 Quisenberry, LCDR M. Ellen, 328
 Ree, Malcolm James, 542
 Reiser, G. A. (Greg), Captain, 704
 Rheinstein, Julie, 298
 Ricotta, Frank J., Jr., 189
 Riedel, Sharon L., 183

Rivkin, David W., 621
 Root, James T., 876, 901
 Rose, Andrew M., 627
 Rosenbaum, Harvey, 399
 Ross, Peter J., Lieutenant Commander,
 Royal Navy, 692
 Ross, William A., 888
 Rossmeyssl, Paul G., 45
 Ruck, Hendrick W., Dr., x, 304, 310,
 353, 809
 Rucker, Linda, M.A., 129
 Rueter, F. H., 797
 Russell, Craig J., 200
 Saia, F. Edward, 177
 Salter, Charles A., 515
 Salvato, Annette M., 51
 Samet, Michael G., 492
 Sands, W. A., 458
 Sarli, Gary G., 767
 Schambach, Sibylle B., 111
 Schlichting, Christine L., 70
 Sessions, G. Rufus, 504
 Sharf, James C., 428
 Shaw, Fred, 256
 Short, Lawrence O., Lt Col, 262
 Shukitt, Barbara L., 76
 Shukitt, B. L., B.A., 76, 440
 Siebold, Guy L., 87
 Siem, Frederick M., 565
 Simutis, Zita M., 810
 Smith, Alfred L., Jr., 231
 Smith, Elizabeth P., 647
 Smith, Norman D., 9, 521
 Smootz, Edwin R., 15
 Sorensen, H. Barbara, Dr., 722
 Sotello, Wendy L., 1Lt, USAF, 378
 Sprenger, William D. Ph.D., 183, 486
 Staley, Michael R., 292, 241
 Stanley, Paul P., II, 359, 372, 378
 Steege, Friedrich W., 117
 Steinberg, Alma G., 171
 Stewart, John E., II, 21
 Stone, Leroy, Ph.D., ABPP, ABFP, 416
 Stricker, Lawrence J., 194
 Sullivan, Edward W., Major, 153
 Sulzen, Robert H., 900, 919

Swartz, Merryanna, 832
 Symington, Lawrence E., 51
 Tartell, J. S., x, 305, 335, 353, 372
 Taylor, Barbara, 663
 Teitelbaum, Joel M., Ph.D., 237
 Terry, Marijane, 256
 Terry, Pamela M., 761
 Thorne, David R., 504
 Thorsden, Marvin L., 153
 Tiffany, John R., 9
 Treichel, Dr. Janet M., 316
 Tremble, Trueman R., 105
 Tucker, Mollie J., 659
 Twohig, Paul T., 105
 Usova, George M., Ph.D., 588
 Van Nostrand, Sally J., 57
 Vandivier, Phillip L., 250
 Vaughan, Paul R., 608
 Vaughan, David S., 779, 803
 Ventura, Jose G., Jr., Major, 906
 Vestewig, Richard W., Ph.D., 843
 Wainer, Howard, 384
 Waldkoetter, Raymond O., 250
 Walker, Martin R., 647
 Walker, Clinton B., 640
 Walker, C. Lee, 602
 Walthers, Michael A., SSG, 440
 Wasaff, Samuel, Colonel, xvii
 Watson, Thomas W., 542
 Weaver, Charles N., 189
 Weeks, Joseph L., 749
 Weissmuller, Johnny J., 292, 335, 360,
 366
 White, Leonard A., 811
 White, Charles R., Brigadier General
 716
 Whitmarsh, Patrick T., 923
 Wiebe, I. E. (Ed), Captain, 704
 Wild, W. R., Major, 99
 Wildgrube, Wolfgang, 446
 Williams, John E. 1Lt, USAF, 378
 Wise, Laurence L., 131
 Wiskoff, Martin F., 210
 Wolfe, John H., 463
 Wynkoop, Keith R., Captain, 16
 Yadrack, R. M., 791